CLUSTERING METHODS FOR MULTI-ASPECT DATA

Doctor of Philosophy

by

KHANH THI NGOC LUONG

Master of Science

Hanoi University of Education, Vietnam

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

School of Electrical Engineering and Computer Science
Science and Engineering Faculty
Queensland University of Technology

2019

Keywords: Multi-view Data Clustering/Learning, Multi-type Relational Data Clustering/Learning, Multi-Aspect Data Clustering/Learning, Non-negative Matrix Factorization, Manifold Learning, k nearest neighbour graph, p farthest neighbour graph, Multiplicative Update Rule, Gradient Coordinate Descent.

Statement of Contribution of Co-Authors for Thesis by Published Papers

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

In the case of this thesis, Chapter 2, Chapter 3, Chapter 4 and Chapter 5 are formed based on six papers. Chapter 3 contains Paper 3 and Chapter 5 contains Paper 5, which include the collaboration with other PhD students. The certifying authorship has been addressed in detailed in these corresponding chapters. Other

papers that are included in this thesis are co-authored by the PhD candidate, Khanh Thi Ngoc Luong and the candidate's principle supervisor, Associate Professor Richi Nayak. Apart from that no one else has contributed in these published papers. The candidate and her supervisor have agreed to use these publications as a part of this thesis.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.

Name: Dr Richi Nayak, Associate Professor

Signature: QUT Verified Signature ***Date:*** 24-06-2019

Statement of Original Authorship

In accordance with the requirements of the degree of Doctor of Philosophy in the School of Electrical Engineering and Computer Science
Sience and Engineering Faculty, I present the following thesis entitled,

CLUSTERING METHODS FOR MULTI-ASPECT DATA

Doctor of Philosophy

This work was performed under the supervision of Associate Professor Richi Nayak. I declare that the work submitted in this thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at Queensland University of Technology or any other institution.

KHANH THI NGOC LUONG

Signature: QUT Verified Signature ***Date:*** 24-06-2019

Acknowledgements

I would like to start my acknowledgement by expressing my thankfulness and appreciation to my Principle Supervisor, Associate Professor Richi Nayak, who has strongly and continuously supported me throughout my PhD journey. I would like to thank her for her encouragement, her motivation, and her extensive knowledge. I believe that she has encouraged me to grow as a research scientist.

I gratefully thank my mother, my father, my parents-in-law, my husband and my daughters for their great and unconditional love. Thanks also go to my extended family and my best friends for their encouragement during my PhD.

Special thanks to Ton Duc Thang University for their kind supports throughout my study, my scholarship sponsor, QUT-VIED, for their financial supports, and the Science and Engineering Faculty, QUT, which provides me a comfortable research environment and assistance. I acknowledge the services of professional editor, Diane Kolomeitz, who provided copyediting and proofreading services, according to the guidelines laid out in the University-endorsed national policy guidelines.

Thank you to all the colleagues in my research group and the Data Science discipline for bringing to my study journey various helpful and widely ranging discussions and also for being great friends to me.

Abstract

With the rapid growth of computational technology, multi-aspect data has become ubiquitous. It can be represented by multiple aspects such as multiple types of relationships, called Multi-type Relational Data (MTRD) and multiple types of features, called Multi-view data. In the last decade, multi-aspect data clustering has become a promising research topic due to the embedding of rich information for data learning and returning informative outcomes. The research problem here is to simultaneously cluster many interrelated data types, based on their relationships, into common groups for the MTRD data and to employ all view data to learn these common data groups, with consensus amongst all views for the multi-view data.

Multi-aspect data naturally exhibits high dimensionality and sparseness. The challenge in clustering multi-aspect data is to exploit all consensus and complementary information from multiple aspects. Sophisticated methods are required that can simultaneously and effectively learn all consensus and complementary information for producing accurate and useful clustering outcomes.

Many approaches have been proposed to cluster multi-aspect data with associated strengths and shortcomings. This thesis uses Non-negative Matrix Factorization (NMF) and manifold learning combination as the fundamental framework to deal with the high relatedness, sparseness and high dimension data. The combination of NMF and manifold learning, works as low-rank representation learning from the high-dimensional and sparse data for the clustering task. This approach has shown effectiveness in traditional one-aspect data because of the ability to find the

good low-rank representations embedding the latent features of data and respecting the geometric shape of the original data. However, it requires a sophisticated design to effectively work on multi-aspect data. For example, it should be able to: (1) incorporate many types of relationships; (2) exploit and represent both the compatible and complementary information in the data; and (3) learn the useful and correct manifold, which is always a challenge of using the manifold learning technique, even in traditional one-aspect data. This thesis proposes various learning methods for multi-aspect data to address these challenges.

Firstly, two methods, ARASP (Learning the Association Relationship and Accurate Geometric Structure for Multi-Type Relational Data) and DiMMA (Learning the Diverse Manifold for Multi-Aspect Data), have been developed for MTRD data. In ARASP, more relationship information has been involved in the learning process and an accurate and more informative manifold is learned where all important close and far distances between points have been encoded when learning the manifold for data representation. In DiMMA, a complete and comprehensive manifold is learned for multi-aspect data, by including a proposed inter-manifold term into the learning process. The proposed term is shown to play the role of a ranking term and can work effectively on high-dimensional data. Though ARASP and DiMMA are designed for MTRD, they can be easily customised to be deployed in multi-view data clustering.

Secondly, two methods, MVCMF (Multi-view Coupled Matrix Factorization) and 2CMV (Utilising the Consensus and Complementary Information for Multi-view Data), have been developed for multi-view data. In MVCMF, the coupled matrix factorization (CMF) framework is proposed for multi-view data learning, which shows benefits for learning the consensus low-rank representation for multi-view data. The CMF is also shown to be very efficient in computational complexity, since the number of factor matrices to be updated has been decreased by half as compared to NMF-based methods. In 2CMV, an integration of both NMF and CMF is proposed to utilize both consensus and complementary information in multi-view

data. An optimal manifold, which is the most consensed manifold, is learned in 2CMV, and a novel complementary enhancing constraint, enable the 2CMV to learn comprehensive information for multi-view data learning.

Comprehensive review on state-of-the-art multi-aspect data clustering methods and extensive experiments conducted throughout the thesis, show the effectiveness of the developed methods. An application of ARASP is proposed to the community detection problem showing the ability of multi-aspect data clustering methods. Methods developed in this thesis show that fully exploiting all available relationships and information in the multi-aspect data, as well as effectively learning a more accurate and meaningful manifold, will significantly improve the clustering performance.

In summary, the thesis has contributed to the development of multi-aspect data clustering and the data mining field by successfully dealing with many challenging problems that have arisen in the literature and in real-world applications.

Contents

List of Figures

List of Tables

List of Abbreviations

- MTRD – Multi-type Relational Data
- NMF – Non-negative Matrix Factorization
- MUR – Multiplicative Update Rule
- GCD – Greedy Coordinate Descent
- kNN – k nearest neighbour graph
- pFN – p furthest neighbour graph
- NMI – Normalized Mutual Information
- CMF – Coupled Matrix Factorization
- CD – Coordinate Descent
- HALs – Hierarchical Alternating Least Squares algorithm
- FHALs – Fast Hierarchical Alternating Least Squares algorithm
- ONMTF – Orthogonal Nonnegative Matrix Tri Factorization method
- MVCMF – Coupled Matrix Factorization for Multi-view data method
- GNMF – Graph Regularized Non-negative Matrix Factorization method
- MultiNMF – Multi-View NMF method
- NMTF – non-negative tri factorization framework

- STNMF − Symmetric Non-negative Matrix Tri-Factorization method
- DRCC − Dual Regularized Co-Clustering method
- RMC − Relational multi-manifold co-clustering method
- RHCHME − Robust High-order Co-Clustering via Heterogeneous Manifold Ensemble method
- MMNMF − Multi-Manifold Regularized Nonnegative Matrix Factorization framework
- CBCC − Consistent Bipartite Graph Co-Partitioning method
- MVSC-CEV − Multi-view Spectral Clustering by Common Eigenvectors method
- S-CPC − Stepwise common principle components method
- SRC − Spectral Relational Clustering method
- MVSC − Multi-view subspace clustering method
- MLRSSC − Multi-view Low-rank Sparse Subspace Clustering method
- ARASP − Learning Association Relationship and Accurate Geometric Structures for Multi-type Relational Data
- SPNMF − Structure Preserving Non-negative Matrix Factorization
- ARP − Association Relationship Preserving method
- BiORNM3F − Bi-orthogonal 3-Factor
- MultiOR-NM3F − Multi-Orthogonal 3-Factor
- DiMMA − Diverse Manifold for Multi-Aspect Data method
- MVCMF − Multi-view Coupled Matrix Factorization method

- PLSA − Probabilistic Latent Semantic Analysis
- AMVNMF − Adaptive Multi-View Semi-Supervised Non-negative Matrix Factorization method
- MC-NMF − Multi-component nonnegative matrix factorization
- DiNMF − Diver NMF method
- 2CMV − Utilising Consensus and Complementary information for Multi-view data

Chapter 1

Introduction

This chapter presents the overview of the research including the background, problem, questions, aims and objectives of the research. The overall research significance and limitations are described accordingly. The structure of the thesis is also presented at the end of the chapter.

1.1 Background

Clustering, which aims to extract useful information from unlabelled data through a process of finding natural groupings in the data based on their similarities, is a vital problem in data mining. This research area of unsupervised machine learning has been studied for decades in many fields such as data mining, text mining, image processing, web analysis and bio-informatics. Most existing clustering methods, although showing effectiveness, are designed for traditional homogeneous data where the data sample is represented by one type of feature or is considered under one perspective, called a view. However, most datasets nowadays are getting richer in terms of both semantics and structures that involve multiple types of features and relationships, which are called Multi-type Relational Data (MTRD) or multiple perspectives, which are called Multi-view Data. These datasets are becoming the

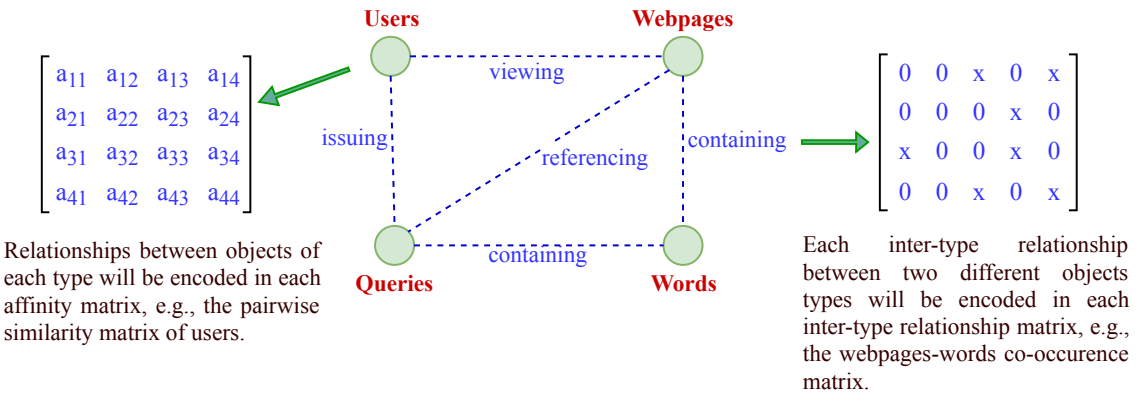


Figure 1.1: Example of Multi-type Relational Data with four object types: Webpages, Words, Users and Queries and various relationships between them.

main focus of most data mining techniques in this decade.

In an MTRD, there exist many types of objects and various relationships among these interrelated objects. Different object types include sample object type and different feature object types. Different relationship types are the **intra-type relationship**, representing relationships between objects of the same type and the **inter-type relationship**, that encodes the relationships between objects of different object types. An example of MTRD is a web search system that can include four kinds of objects namely Webpages, Users, Queries and Words, as well as different kinds of relationships namely issuing (between Users and Queries), viewing (between Users and Webpages), referring (between Queries and Webpages) and containing (between Queries and Words or between Webpages and Words) as illustrated in Figure 1.1.

Examples of multi-view data are given in Figure 1.2 where data can be available under multiple sources (e.g., news can be existing from BBC, Yahoo etc. Figure 1.2.a) or many forms such as text, image as in Figure 1.2.b. The multilingual dataset as shown in Figure 1.2.c is a typical multi-view dataset with each document represented by many different language translations. An image can be described by many different features such as Edge, Fourier or Texture as in Figure 1.2.d. MTRD and multi-view data are becoming ubiquitous nowadays. Both the datasets share

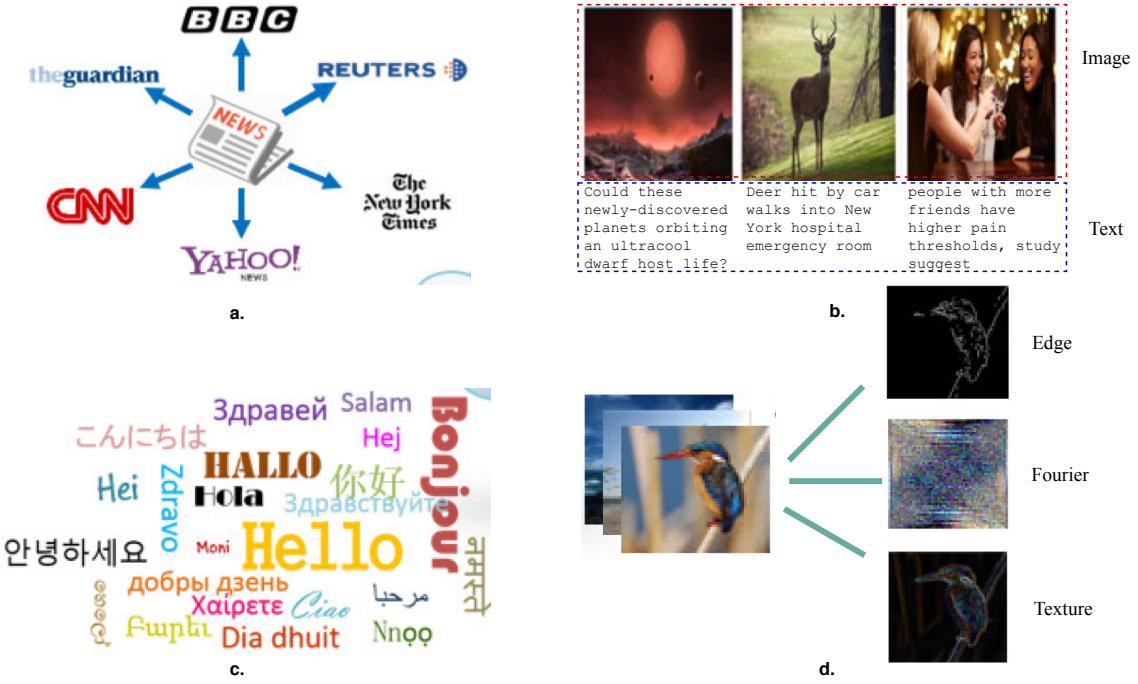


Figure 1.2: Examples of Multi-view Data [101].

the same property of providing complementary information for the learning process, yet show different focuses, i.e., relationship types or views. This thesis will treat these two dataset types as **multi-aspect** data where different aspects in **multi-view** data are different views, and different object types together with their associated relationships are different aspects in **MTRD** data. As a result of embedding rich information, clustering on multi-aspect data will significantly improve performance and quality [101, 120]. However, directly applying traditional one-way clustering methods that too heavily rely on features (or attributes) of the multi-aspect data may lead to unrealistic and unsatisfactory solutions, since they fail to exploit the latent relatedness between samples and different feature objects in MTRD or to exploit the latent property of data on each view representation in multi-view data [70, 99]. This fact leads to the need of designing methods specifically for clustering multi-aspect data in order to identify the underlying structure accurately and extract more useful and meaningful knowledge.

There have been a wide-range of MTRD and multi-view clustering methods developed based on the concepts of graph partitions (spectral clustering) [28, 59, 62,

70, 90], subspace learning [19, 35, 104, 105] and nonnegative matrix factorization (NMF) [48, 56, 99, 103, 120]. These approaches can identify the underlying structures in the multi-aspect data and show clearly improved performance compared to traditional clustering methods. Yet, due to the requirement of taking all possible available information into account to learn meaningful outcomes and the natural complexity of multi-aspect data, clustering on MTRD and multi-view data remains a challenging problem and one to which it is worth paying meticulous attention.

1.2 Problem Statement and Motivation

1.2.1 Problem Statement

Clustering on multi-aspect data is the process of using all possible information to explore the dependence and latent relationships amongst all aspects to form common groups in the data. For MTRD, the clustering problem is considered as a process to cluster several related data types simultaneously by effectively exploring all possible intrinsic relationships to obtain the true latent structures in data [70]. The simultaneous processing of all types and relations enables the clustering analysis to consider all object types and their correlations to have a general look on data and helps to yield an accurate and meaningful clustering solution. For multi-view data, the task is how to exploit all available views' data, to learn the successful outcome that embeds the complementary and compatible information from all views [107, 117]. In multi-view data learning, the focus is not to utilize all the relationships as in MTRD but to utilize the consensus and complementary information embedded from all views. This thesis will mainly focus on clustering these two dataset types and aims to effectively yield the meaningful clusters by using all valuable supplementary information and correlations between data.

1.2.2 Motivation

The multi-aspect data is normally very sparse and high dimensional [108, 110]. Concatenation of data existed in multiple aspects in a single matrix, will fail to include the interrelatedness between data and cause the low performance. This fact leads to most of the existing state-of-the-art methods using NMF framework and manifold learning as the dimensionality reduction technique to simultaneously learn the true embedded low-dimensional structures from multiple aspects before conducting the clustering task. NMF [63, 64] seeks decomposition of the original matrix into two low-order matrices. NMF has been proved to be able to learn the part-based representation, therefore it has been widely used in aiding the clustering task by finding the low-rank representation that embeds the original latent structure [63]. Manifold learning [10] is long known for non-linear dimensionality reduction by preserving the local geometric structure of original data to the projected space via a process of ensuring the distances between points within a neighbourhood area using the k nearest neighbour (kNN) graph. Prior research [21, 43, 48] has shown that the combination of NMF and manifold learning algorithms, which helps in learning and ensuring the geometric structure of data when projecting to low dimensional space, will help in learning the good representation respecting the original data structure for the clustering task. However, utilising NMF and manifold learning in multi-aspect data is a challenging problem that has attracted much attention in recent years.

The NMF framework is flexible and can do both hard and soft clustering thus it is convenient to be extended to the multi-aspect data problem, however, application of the NMF framework on MTRD and multi-view brings different difficulties. For MTRD data, the challenge is to make use of all possible valuable relationships in an effective manner, despite the natural complexity of the data. For multi-view data learning, the objective is to exploit all available views' data to learn the success-

ful outcome that embeds the complementary and compatible information from all views. A common approach is to separately learn the low-rank representations from all views and find the consensus representation in the fusion step [56, 58, 102, 103]. However, the late fusion step may lead to the degradation of the clustering performance since they disregard the contributions of different views to the learning process. The thesis will develop methods based on the NMF framework to effectively learn the low-rank representations for both MTRD and multi-view data in the manner that all important relationship information will be utilized in MTRD and the consensus latent structure in multi-view data will be learned naturally during the learning process.

Though manifold learning has been proved to help clustering to achieve the more meaningful clusters as a result of the ability to ensure the local geometric structure of data unchanged during the high to low order mapping process, however, manifold learning also brings other difficulties to the problem. The manifold learning due to relying on a kNN graph has to deal with the problem of choosing the neighbourhood size which is still daunting. The application of manifold learning on multi-aspect data is not a trivial extension due to the complexity of multiple aspects. This thesis will deal with these challenges and aim to propose methods, taking advantage of the NMF framework and manifold learning in an effective manner to learn the meaningful outcomes for multi-aspect data.

1.3 Research Questions

Clustering on multi-aspect data results in purposeful outcomes however the complexity of multi-aspect data poses several challenges to the learning process. The high-level question that this thesis aims to address is: how to develop clustering methods that can effectively work on multi-aspect data by utilising all available information? Prior research on multi-aspect data clustering shows that the multi-

aspect data is usually sparse and high dimensional [108, 110]. The NMF and manifold learning group of methods have been shown to deal with the high dimensional and sparse data effectively [21]. The NMF and manifold learning-based methods for multi-aspect data clustering will be the focus of this thesis. This thesis will explore how to effectively apply NMF and manifold learning for MTRD and multi-view data. More specifically, the detailed questions are listed as below.

Question 1. Data – What are the distinct characteristics of MTRD and multi-view data as compared to the traditional data that can affect the performance of clustering?

a. What are the identifying characteristics of multi-aspect data?

b. What are the main focuses and/or the expected outcomes of clustering on each dataset type?

Question 2. Methods – Based on the characteristics of multi-aspect data, how to apply the NMF framework and manifold learning on multi-aspect data successfully?

a. Why is the NMF and manifold learning-based framework selected?

b. How is the NMF and manifold learning-based framework deployed on multi-aspect data?

c. What are the challenges of incorporating manifold learning algorithms in the NMF framework on multi-aspect data where there are various relationships and many types of feature are represented?

d. How can an NMF and manifold learning-based method achieve the high clustering outcome on multi-aspect data?

Question 3. Relationships – What types of relationships exist in multi-aspect data? How to successfully make use of them?

a. What types of relationships exist in multi-aspect data?

b. What are the challenges to make use of these relationships?

c. What benefits do the diverse relationships bring to the learning process?

Question 4. Constraints – What are some significant constraints that have been used in clustering and how they can be utilized in multi-aspect data for an improved clustering performance?

This question arises from the fact that constraints play important roles in the learning process as evidenced in many works [29, 48]. Finding the answer to how to productively use and handle different types of constraints in the appropriate situation will significantly boost the clustering performance.

Question 5. Application – Can the developed clustering methods be deployed in closely related real-world problems?

This is last but not the least question. Can the multi-aspect data clustering methods developed in this thesis be applied in other related problems such as the community discovery problem in a social network that aims to seek for the latent groups?

1.4 Research Aims and Objectives

The main aim of the proposed research is to build effective clustering methods that can result in the meaningful clusters for multi-aspect data especially MTRD and multi-view data. The objectives of this research are listed in detail as follows:

RO.1. Developing MTRD clustering methods.

This thesis aims to investigate the limitation of the NMF framework and the limitation of application of the existing NMF-based methods for MTRD. More specifically, the limitation of the NMF framework is the failure of preserving the local geometric structure of original data; and the limitation of existing MTRD clustering methods

based on NMF framework is the ability to exploit all available information in the data.

- **RO.1.1.** Though many methods have been developed to deal with this limitation, how to learn accurate and useful geometric structure of original data for effective learning is still challenging.

- **RO.1.2.** An urgent requirement is how to effectively utilize inter-type relationships as well as intra-type relationships in MTRD. Other types of relationships exist, such as association relationships between clusters of different object types that have not been considered.

Developing the MTRD clustering methods responding to these challenges will be the major objective of the thesis.

RO.2. Developing multi-view clustering methods.

The main focus of a multi-view clustering method is to learn the compatible and complementary information existing in multiple views. The thesis will investigate the advantages and disadvantages of using the NMF framework in pursuing the goal of multi-view data clustering.

The NMF framework applied in multi-view data is normally conducted separately on each view to learn the low-rank representation before calculating the main consensus low-rank representation. This may lead to the loss of a natural associative relationship among views. The combination of NMF and manifold learning for multi-view data should be investigated accordingly. Developing the appropriate factorizing-based framework with the aid of manifold learning, to ensure that the learning process can effectively utilize both consensus and complementary information presented in multi views, will be the key objective of the thesis.

RO.3. Effectively and flexibly utilising different constraints.

When developing multi-aspect clustering methods, most available types of constraints will be studied to investigate their applications in an effective manner. There are many cases of applying constraints (can also be referred as regularizers or smoothing methods) that show advantages but also pose challenges. For example, l_1-norm is good for representative purposes but causes sparseness. Orthogonality brings unique and elegant solutions but is too strict, which leads to slow convergence. Non-negativity is useful in reality but cannot be used in some cases. l_2-norm brings an even distribution for data samples on all groups however it does not always show effectiveness. This thesis will thoroughly study these constraints. It is expected that the outcome will bring much valuable and interesting information for unsupervised learning.

RO.4. Applying the proposed method to a real-world problem.

Clustering techniques have been found to be fairly close to the problem of discovering communities [5]. A specific objective is to apply a multi-aspect approach to the community detection problem where it can effectively utilize all existing relationship information such as the structural relationship information or the content relationship information in a Twitter dataset in order to produce the worthwhile user groups. The success of the application will bring great potential to the multi-aspect learning for other real-world problems.

1.5 Research Contributions

This thesis has developed several methods of clustering multi-aspect data. The contributions are listed as below.

RC.1. Provide a comprehensive survey

As the first contribution, the thesis has conducted a comprehensive survey on the problem of clustering MTRD and multi-view data, an under-studied problem. Three groups of methods - spectral-based group, subspace learning and NMF-based group - have been reviewed. Since the thesis focuses on NMF-based methods, more attention has been paid to this group. Extensive experiments on the state-of-art methods belonging to this group have been conducted to investigate the effectiveness of methods on diverse datasets. This has led to a chapter published in a Springer book titled *Linking and Mining Heterogeneous and Multi-view Data*, 2019.

RC.2. Develop MTRD Clustering Methods

- *RC.2.1. Learning Association Relationship and Accurate Geometric Structure for MTRD (ARASP) method*

  In this method, a novel fashion to build the affinity matrix for each object type is proposed for the MTRD dataset in order to include the far distance information of objects in manifold learning. This affinity matrix is used to learn and preserve the accurate manifold when projecting data from high to low dimensional space. This novel construction of the affinity matrix, embedding both low and high relatedness, can also be used in dimensionality reduction and on other traditional data learning. Another contribution of the proposed ARASP is the association relationship between clusters of different object types to be embedded into the NMF framework through enforcing the Normalized Cut type Constraint. The incorporation of an association relationship boosts the learning process to be faster to achieve the optimal solution and bring the consequential solution for MTRD clustering. The result of this research is published as the tier A* conference paper in the 34th International Conference on Data Engineering (ICDE 2018).

- *RC.2.2. Learning Diverse Manifold for Multi-Aspect data (DiMMA) method*

  Inspired from the fact that most NMF and manifold learning-based clustering methods on MTRD care for the manifold of each data type using intra-type relationship only, this may lead to the incompleteness of information in the manifold learned from MTRD. DiMMA is proposed to learn a diverse and more complete manifold for MTRD by constructing and preserving manifold generating from both the intra- and inter-relationships. The manifold learned with this manner is believed to be useful and complete, thus more informative results can be achieved. It is observed that the proposed inter-manifold learning is able to aid in the learning process on high-dimensional data as evidenced by empirical analysis. This research has resulted in a manuscript, which has been submitted to IEEE Transactions on Knowledge and Data Engineering (TKDE) Journal and it is now under the major revision process.

RC.3. Develop Multi-View Clustering Methods

- *RC.3.1. Coupled Matrix Factorization for Multi-view data (MVCMF) method*

  MVCMF presents a novel unified coupled matrix factorization framework with the complementary similarity information to effectively learn the consensus representation for multi-view data. MVCMF also proposes to use Greedy Coordinate Descent (GCD) instead of Multiplicative Update Rule (MUR) as the new optimizing scheme for the coupled matrix framework. This work has led to a conference paper, published in the 19th International Conference on Web Information Systems Engineering (WISE 2018), a tier A conference.

- *RC.3.2. Exploiting the Consensus and Complementary Information for Multi-view data Represenetation and Learning (2CMV) method*

  The 2CMV aims to utilize both consensus and complementary information for effective data analysis. A factorization-based loss function to learn two compo-

nents encoding the consensus and complementary information for multi-view data by using Non-negative Matrix Factorization (NMF) and Coupled Matrix Factorization (CMF) has been designed. Furthermore, an optimal manifold which embeds the common latent structure of multi-view data is proposed in this method. More importantly, a newly complementary enhancing term is included in the loss function to be able to bring more comprehensive information for a more insightful clustering outcome. This research has resulted in a paper, which has been submitted to the 21st International Conference on Computer Vision (ICCV) 2019, a top-tier conference.

RC4. Application in a real-world related problem

An application of the ARASP method in community discovery is investigated where the Twitter data is used to create different MTRD datasets with the content and structure information. Experiment results showed the effectiveness of approaching the community discovery under the MTRD context as well as the ability of using a multi-aspect clustering method in an important real-world problem. This work has been accepted to be published in the 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019).

1.6 Published Papers

The list of published papers as part of this research is given below,

Paper 1. **Khanh Luong**, Richi Nayak (2019), *Clustering Multi-view Data Using Non-negative Matrix Factorization and Manifold Learning for Effective Understanding: A survey paper.* Linking and Mining Heterogeneous and Multi-view Data, Springer Publisher. (Will form part of Chapter 2)

Paper 2. **Khanh Luong**, Richi Nayak (2018), *Learning association relationship and accurate geometric structures for multi-type relational data.* In the 34th International Conference on Data Engineering (ICDE 2018), 16-19 April 2018, Paris, France. (Will form part of Chapter 3)

Paper 3. T.M.G, Tennakoon, **Khanh Luong**, Wathsala Mohotti, Sharma Chakravarthy, and Richi Nayak, *Multi-type Relational Data Clustering for Community Detection by Exploiting Content and Structure Information in Social Networks.* The 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019) (In Press). (Will form part of Chapter 3)

Paper 4. **Khanh Luong**, Richi Nayak, *Learning the Diverse Manifold for Meaningful Multi-Aspect Data Representation.* IEEE Transactions on Knowledge and Data Engineering (TKDE) Journal (Under major revision). (Will form Chapter 4)

Paper 5. **Khanh Luong**, Thirunavukarasu Balasubramaniam and Richi Nayak (2018), *A novel technique of using Coupled Matrix Factorization and Greedy Coordinate Descent for Multi-view Data Representation.* Lecture Notes in Computer Science: The 19th International Conference on Web Information Systems Engineering – WISE 2018. (Will form part of Chapter 5)

Paper 6. **Khanh Luong**, Richi Nayak, *Learning the Consensus and Complementary Information for Multi-View Data Clustering.* In 21st International Conference on Computer Vision (ICCV 2019) (Under review). (Will form part of Chapter 5)

1.7 High level overview of the thesis

This section will outline the high-level relationships between the six published/under-review papers in the thesis.

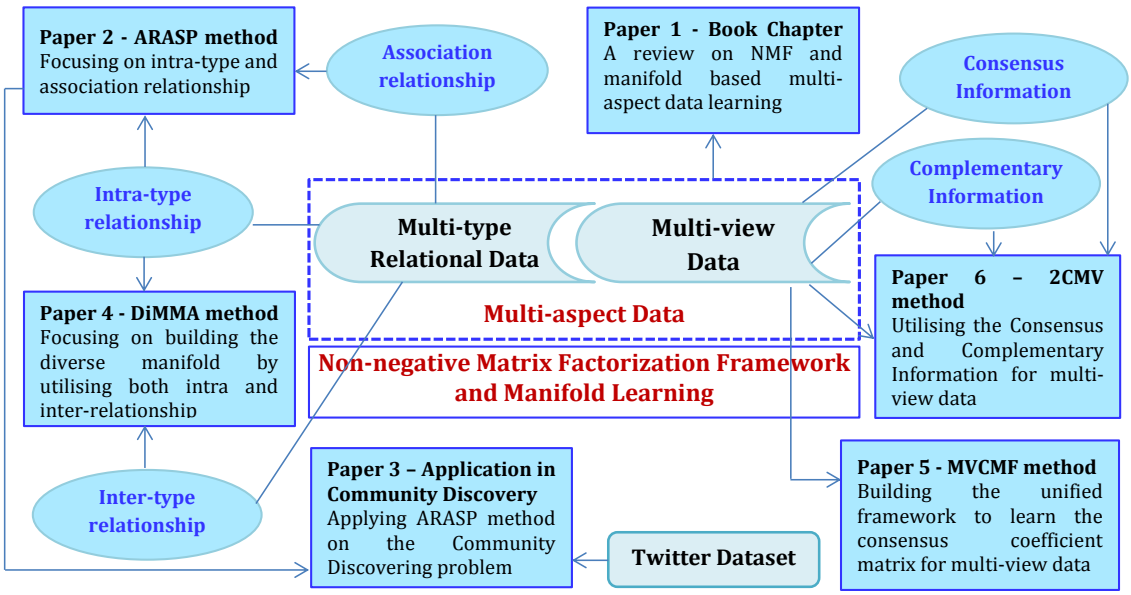


Figure 1.3: High level overview of the thesis

As can be seen from Figure 1.3, there are two dataset types being dealt with in this thesis, i.e., MTRD and Multi-view data. All methods proposed in this thesis are based on NMF framework and manifold learning to be effective for multi-aspect data, which is normally high-dimensional and sparse. All important components such as the inter- and intra-relationships, the association relationship in MTRD data and the important principles of learning the consensus and complementary low-rank representation in multi-view data, have been effectively exploited in these methods.

Firstly, a comprehensive theoretical and empirical review on the group of methods combining NMF and manifold learning for multi-aspect data has been conducted. This has resulted in Paper 1.

Secondly, the ARASP (Learning Association Relationship and Accurate Geometric Structure for MTRD) method is proposed in Paper 2 that aims to learn the

accurate manifold by using extensive intra-type relationships based on both close and far distances neighbours. By proposing a novel fashion of building an affinity matrix for each object type, ARASP reduces the dependence of choosing k to find k-NN neighbours. Incorrect neighbourhood size may lead to an incorrect manifold learned, thus it can decrease the clustering performance. The association relationship between clusters of different object types is also embedded in ARASP to obtain the best clustering performance. Paper 3 presents the problem of community discovery as an MTRD problem, where different factors to define the affinities between users can be exploited to form different types of relationships. Paper 3 proposes a novel solution using the ARASP method to obtain the meaningful community groups.

Thirdly, exploiting the inter-type relationship in MTRD, the DiMMA (Learning Diverse Manifold for Multi-Aspect Data) method is proposed in Paper 4 to learn a diverse manifold for the MTRD data by embedding the inter-manifold in the manifold learning process. DiMMA has also reported a promising solution for the high-dimensional multi-aspect data.

Fourthly, the MVCMF (Multi-view Coupled Matrix Factorization) method in Paper 5 is proposed; this aims to naturally learn the embedded consensus low-rank representation from all view data by using coupled matrix factorization in an effort to ensure learning the compatibility in multi-view data. The affinity matrices encoding the pairwise affinities between data samples in each view are incorporated in the method leading to a unified framework of learning low-rank representation for multi-view data.

Lastly, the 2CMV (Learning Consensus and Complementary Information for Multi-View data) method in Paper 6 aims to exploit both consensus and complementary information for multi-view data representation, by using both Non-negative Matrix Factorization and Coupled Matrix Factorization with the aid of learning an optimal manifold from data.

1.8 Research Significance

The popularity of multi-aspect data can be seen in practice by the increased computational resources. It carries rich information for the learning process to mining more valuable knowledge. The developed techniques under the multi-aspect approach contributes to various domains and applications.

First, the thesis has advanced the multi-aspect data clustering and its learning methods by providing many novel methods for multi-aspect data learning. These methods bring new perspectives to the problem of multi-aspect data learning. Typically, (1) they provide the evidence for the effectiveness of the factorization and manifold learning-based technique; (2) They show a new direction of using new manners of factorization-based for multi-view learning which are using coupled matrix factorization or integration of both NMF and CMF; (3) They show the importance of learning more accurate manifolds, by incorporating both close and far distances; by combining both intra- and inter-type distances; or by learning the optimal manifold for multi-view data, to be preserved when projecting to lower order dimensional space.

Second, the thesis also contributes to the problem of dimensionality reduction. Every method designed in the thesis can be adjusted to be applied as a dimensionality reduction algorithm to be further processed by other mining tasks.

Third, the methods developed in this thesis can be used in a wide range of multi-aspect data under the form of MTRD or multi-view data as well as the popular traditional one-aspect data. They are able to applied on many real-world datasets including text, image, sound, video or in bio-informatics data, in diverse problems such as clustering, classification or anomaly detection.

Last but not least, the multi-aspect data learning approach in this thesis can be able to deployed in other real-world problems such as community discovery or collaborative filtering with the novel multi-aspect outlook. The practice of more

informative data being generated requires the learning method to be able to exploit, as much as possible, all the available data in order to capture the true valuable hidden knowledge from data.

1.9 Thesis Outline – Thesis by publication

This is a thesis by publication. The thesis is formed with the following six chapters.

Chapter 1 provides a general overview of the thesis, including research questions, objectives, and significance.

Chapter 2 reviews the problem of MTRD and multi-view clustering in the literature. The chapter will be divided into many sections ranging from the definition of MTRD and multi-view data, and how the two multi-aspect data are related to each other, to different categories of clustering methods. The review will primarily be focused on the group of methods using NMF and manifold learning for MTRD and multi-view data. A list of research gaps will conclude the chapter and play the role of the starting points for development of other methods found in other chapters. This chapter is formed by Paper 1.

Chapter 3 focuses on dealing with the problem of learning the accurate manifold for MTRD together with learning and embedding the association relationship for MTRD. The proposed ARASP method in Paper 2 is applied on a Twitter dataset for the problem of discovering communities and has successfully led to Paper 3. The two papers have formed this chapter.

Chapter 4 works on the context of MTRD and focuses on learning the diverse manifold for the data where the inter manifold is proposed to learn together with the intra manifold. Theoretical analysis on the optimization scheme and convergence are also provided in this chapter. This chapter is formed by Paper 4.

Chapter 5 concentrates on developing methods for multi-view data. Two multi-view clustering methods will be developed in this chapter, focusing on learning

the optimal latent feature for multi-view data. While the first method emphasizes learning the consensus low-rank representation, the second method strengthens on learning both consensus and complementary information from multiple views. This chapter is formed by Paper 5 and Paper 6.

Chapter 6 summarizes the thesis; the significant results and findings of this thesis are aligned with research objectives and research gaps from Chapters 1 and 2. It concludes with recommendations for future research directions.

The breakdown of paper relevance to chapter is given in Table 1.1.

Table 1.1: Thesis outline

Thesis chapter	**Content**
Chapter 1	Introduction
Chapter 2	Paper 1
Chapter 3	Paper 2 and Paper 3
Chapter 4	Paper 4
Chapter 5	Paper 5 and Paper 6
Chapter 6	Conclusion

Chapter 2

Literature Review

This chapter provides an overview of the current literature on multi-aspect data clustering, utilising learning the low-rank representations. It is mainly made up by Paper 1. However, for the purpose of providing a systematic background and context for the whole thesis, Sections 2.1, 2.5 and 2.6 have been added in this chapter. Other sections contain detail of Paper 1. More specifically, the chapter is organized as follow. The first part of the literature review (Section 2.1) presents the related basic concepts of multi-aspect data. The primary focus of this thesis is on the group of multi-aspect learning methods using NMF, where the manifold learning is incorporated to achieve meaningful multi-view representation respecting the embedded geometric structure of data. This family of methods will be discussed in detail in the next part of the chapter and forms the book chapter (Paper 1). Section 2.2 presents the NMF and manifold learning technique used in traditional one-aspect data. Sections 2.3 and 2.4 will discuss in detail the family of multi-aspect data clustering methods based on NMF and based on the combination of NMF and manifold learning. Next, other multi-aspect data clustering approaches such as spectral clustering, subspace clustering and feature concatenation-based clustering, are investigated. A brief review of all these groups of methods is reported in Section

2.5. Section 2.6 will conclude this chapter by highlighting the research gaps, with regard to the main concepts mentioned above.

Paper 1. Khanh Luong and Richi Nayak (2019), Clustering Multi-view Data Using Non-negative Matrix Factorization and Manifold Learning for Effective Understanding: A survey paper. Linking and Mining Heterogeneous and Multi-view Data, Springer Publisher.

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors:

Khanh Luong (PhD Candidate): Conceived the idea, designed and conducted experiments, analysed data, wrote the paper and addressed reviewers comments to improve the quality of paper.

Signature: ***Date:*** 24-06-2019

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Signature: ***Date:*** 24-06-2019

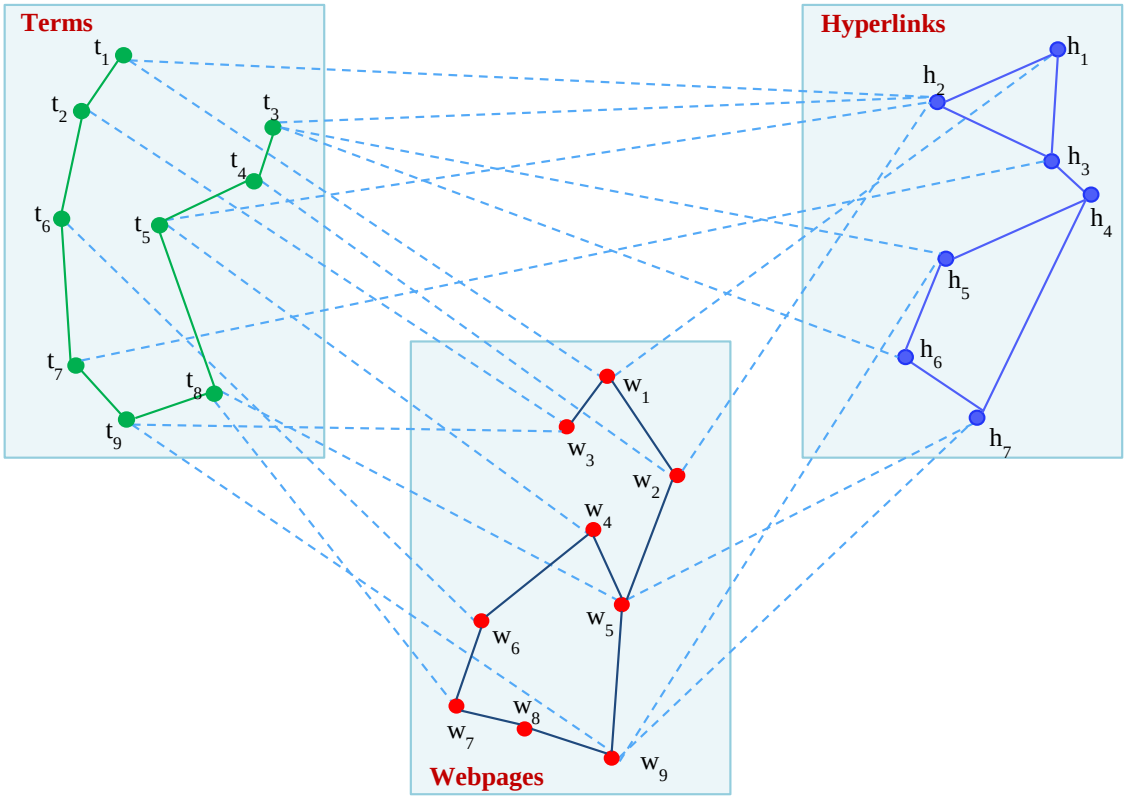


Figure 2.1: Examples of Multi-type Relational Data with three object types: Webpages, Terms, Hyperlinks. The intra-type relationships are represented as solid lines and inter-type relationships are represented as dotted lines.

2.1 Multi-Aspect Learning: Basic Concepts

The multi-aspect data, as noted before, can be (1) multi-type relational data (MTRD) where data exist under various object types and exhibit different relationships among them or (2) multi-view data where data samples are represented via multiple views. The multi-aspect data are becoming ubiquitous and are playing a vital role in practice. This section will provide the details of all important concepts with regards to the MTRD and the multi-view data used in this thesis.

2.1.1 Multi-type Relational Data

The MTRD dataset includes different types of objects as well as different types of relationships. Each aspect can be each object type in the dataset together with the

associated relationships between the object types. For example, the MTRD dataset given in Figure 2.1 has three aspects; the first aspect is the Webpages object type and the relationships between Webpages with other object types.

2.1.1.1 Object type, objects

An object type is a collection of several objects of the same data type. For example, the Webpages object type includes several webpage objects, e.g., w1, w2,... in the MTRD web-search system dataset given in Figure 2.1. Different object types in an MTRD include sample object type and different feature object types. In the given MTRD example, there are three different object types corresponding to three different aspects, which are Webpages, Hyperlinks and Terms. The Webpages object type is treated as **sample object type**; Hyperlinks and Terms are called **feature object types**.

2.1.1.2 Relationship

There are three types of relationships in an MTRD data: inter-type, intra-type and association relationships.

Inter-type relationship, Intra-type relationship

The inter-type relationship in an MTRD describes the relationships between objects of two different object types. The intra-type relationship models the relationships between objects of the same type. In the example in Figure 2.1, inter-type relationships are the relationships between objects of Webpages and Terms, between objects of Webpages and Hyperlinks, and between objects of Terms and Hyperlinks. They are modelled as dashed lines. The intra-type relationships are the relationships between objects within Webpages, between objects within Hyperlinks or between objects within Terms. They are modelled as solid lines. Each inter-type or intra-type relationship is encoded in a matrix. These matrices will be input to the MTRD learning process.

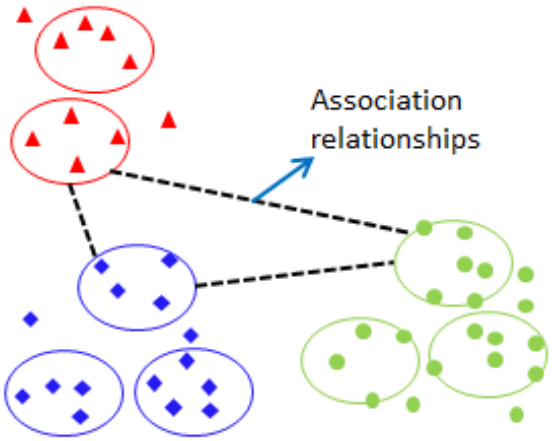


Figure 2.2: Association relationship

Association relationship

The concept of association relationship between clusters of different object types was first proposed in [70] for MTRD. An illustration of association relationships between clusters is shown in Figure 2.2 where it formulates the relations between different clusters of different object types. In this figure, let the red colour clusters be document clusters and the green colour clusters be the term clusters, an association relationship can exist between a document cluster and a term cluster as shown by dashed line in the figure. Learning and controlling the association relationship or the interactions between clusters when learning higher order to lower order mapping will enhance the clustering performance. Inclusion of this type of relationship in MTRD clustering is not much explored and needs more attention.

2.1.2 Multi-view Data

A multi-view data can be seen as a multi-aspect data where each view plays the role of each aspect. In multi-view data, data samples are represented by multiple views (also called multiple features). For example, data samples in the Web search system in Figure 2.3 are Webpages. They are represented by two types of features, i.e., Terms and Hyperlinks, generating two view data, illustrated by the two dotted large rectangles in Figure 2.3. With regard to the relationship concept in multi-view

data, one may notice the relationships between samples on each view (sample-sample relationship denoted by solid lines in Figure 2.3) which have the same meaning with intra-type relationships on a sample object type of MTRD; and the relationships between samples and features (sample-feature relationship denoted by the dashed lines in Figure 2.3), which exhibit the same semantic meaning with the inter-type relationship in MTRD. However, multi-view data may ignore the relationship between different feature types of different views as well as the feature-feature relationships between feature objects on each view. It can be logically deduced that the MTRD exhibits more information, however the multi-view data is simpler and can be easier to handle.

2.1.3 Relationship between MTRD and Multi-view Data

Both MTRD and multi-view data provide the **complementary** information to the learning process as compared to the traditional one-aspect data. Each aspect in multi-aspect data should provide complementary information to the multi-aspect data learning and helps it to exploit the comprehensive representations and meaningful clusters [107]. Though MTRD and multi-view data are similar, they differ in using concepts and have different concerns.

Firstly, apart from complementary constraint, multi-view data requires the **compatibility**, i.e., different data views should embed the consensus latent structure. This principle enforces the cluster structures among all views to go beyond the consensus structure. MTRD does not require the compatibility constraint, since its process of simultaneously clustering all object types using the interrelatedness between object types is corresponding to the process of boosting the clustering performance of sample objects using feature objects.

Second, the focus of two dataset types are different. While MTRD focuses on relationships, multiview data focuses on views and the consensus data learned.

Thirdly, the concept of relationships is different in multi−view and MTRD. The

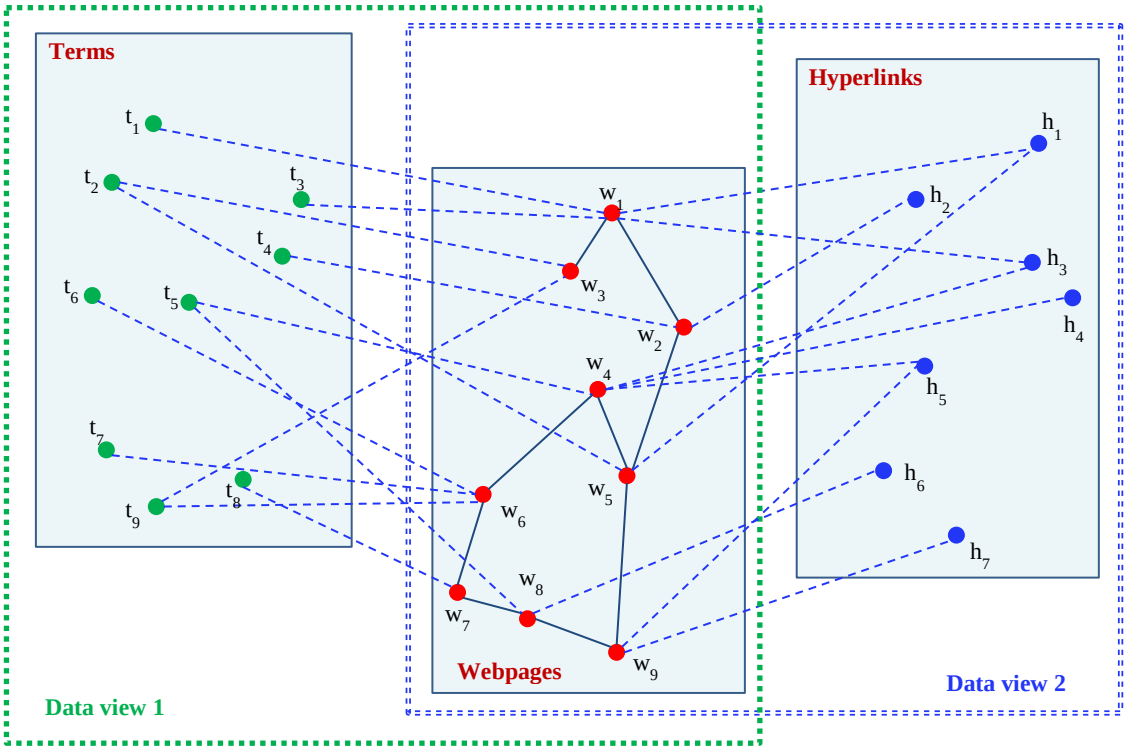


Figure 2.3: Examples of Multi-view Data. The Multi-view dataset is represented by two views Terms and Hyperlinks.

two most well−known types of relationship in MTRD are inter−type and intra−type relationships. For multi−view data, the concept relationship may not be the focus, however, one may expect the sample−sample or sample−feature relationships as discussed in the previous section. The MTRD has association relationship, which means the relationship between different clusters of different object types. For multi−view data, the association relationship can be defined between clusters generated from different views. To the best of our knowledge, there is no method explicitly using this relationship to enhance clustering performance in multi-view data in literature.

2.1.4 Challenge of Learning Multi-Aspect Data

The multi-aspect learning is benefited by using the compatible and complementary information that results in meaningful clusters and higher clustering performance. However, due to being represented with multiple aspects, the underlying multi-aspect data model is usually very sparse and high dimensional. Finding clusters

on the original search space is expensive and results in an unrealistic and unstable clustering solution, since the concept of distance becomes meaningless in high-dimensional space [13]. Therefore, most existing multi-aspect learning methods attempt to learn the low-dimensional latent features of data as a first step for further processing. These methods are based on a non-negative matrix factorization (NMF) framework [48, 99, 117] or subspace learning [79, 82]. Between these two approaches, the multi-aspect clustering methods that utilize NMF have received much attention and have shown to be more effective due to the ability of transforming features and learning the part-based representations that are particularly suitable for clustering task [4, 63, 64]. Subspace learning based methods learn the part-based representations by selecting a subspace of high-dimension. They do not support the features transformation from higher to lower order, therefore are over-performed by NMF.

The next section will review the NMF framework on traditional one-aspect data as well as how it is incorporated with a manifold learning technique to ensure learning the low-rank representation respecting the intrinsic geometric structure of the original data.

2.2 NMF and Manifold Learning Based Clustering Methods on Traditional Data

2.2.1 NMF Based Clustering Methods on Traditional Data

NMF is an established method to find an approximate product of smaller matrices equivalent to the original data matrix for learning lower rank matrices [21, 64]. In clustering, an NMF-based method attempts to project data into the new lower dimension space and then seeks the latent groups in the new low embedded space. Due to the ability to learn part-based representations, NMF has been shown to be

the most effective and flexible clustering method, especially in very high-dimensional data [109]. In the first NMF-based clustering method [109], NMF derived the latent semantic space and determined the document groups based on how each document is combined with each document topic cluster. This first method deploying NMF to document clustering marked a new era. Another milestone work of Orthogonal Non-negative Matrix Tri Factorization (ONMTF) [29], applying the orthogonal constraint on factor matrices of both rows (samples) and columns (features), leads to both non-negative and orthogonal constraints used simultaneously in the NMF framework. These two constraints on factor matrices turn these matrices into cluster indicator matrices, thus making ONMTF a well-defined NMF co-clustering based method. These works established that an NMF-based clustering solution is not only unique but it is also meaningful and interpretable. NMF-based clustering methods have shown the effectiveness of NMF over other traditional clustering methods such as K-means or hierarchical clustering methods [8, 60]. The NMF framework formulation is provided as below,

Given a data matrix $X \in R^{n \times m}$ representing n rows of data samples with m features, NMF will find non-negative factors $H \in R^{n \times k}$, $W \in R^{k \times m}$ of X such that the product of H and W will be an approximation of X. The NMF framework aims to minimize the following objective function,

$$\min \|X - HW\|_F^2, \text{ s.t., } H \geq 0, W \geq 0 \tag{2.1}$$

Factor matrix H learned from Eq. (2.1) is considered as the low-rank latent feature matrix on the newly mapped space learned from data matrix X.

NMF-based methods relying on the loss function in Eq. (2.1) focus on finding the approximate factor matrices only, and the framework has been criticized to fail to preserve the geometric structure of data [21]. In other words, the geometric structure of the learned low-order representation may not share the similar geometric structure

as the original structure of the original data, which may lead to an unrealistic clustering result.

Next presented is the manifold learning technique [11], which has been known to learn the intrinsic manifold or geometric structure of data and how to incorporate it into the NMF-based methods.

2.2.2 NMF and Manifold Learning Based Clustering methods on Traditional Data

Manifold learning is the process to learn and preserve the intrinsic geometric structure embedded in the high-dimensional data [11, 96]. These algorithms work on the assumption that the nearby points should have similar embedding. This assumption requires the learning process to ensure the distances between points remain unchanged. There exist many manifold learning algorithms. Some algorithms, e.g., MDS [27], ISOMAP [97] attempt to learn and preserve the distances between all data points. These methods require expensive computational complexity as well as a clustering task that does not need all distance information to be preserved [97]. Therefore, when applying on clustering, the manifold learning algorithms to learn and preserve the local geometric structure of data points [11] are preferred. This manifold learning algorithm [11] presents the following optimizing term to ensure the data points in the new space are smooth, with the intrinsic geometric structure of the original data.

$$\min \sum_{i,j=1}^{n} \|h_i - h_j\|^2 a_{ij} \tag{2.2}$$

where $\|h_i - h_j\|^2$ is the Euclidean distance estimating the closeness between two new representations h_i and h_j projected for arbitrary x_i, x_j data points. $A = \{a_{ij}\}^{n \times n}$ is the adjacency matrix captured by building the k nearest neighbour (kNN) graph,

a_{ij} is defined as [11, 21],

$$a(i,j) = \begin{cases} t_{ij} \text{ if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{2.3}$$

where $T = \{t_{ij}\}$ represents the input similarity information between all pairs of data points x_i and x_j, e.g., the similarities between pairs of documents in a text dataset. $\mathcal{N}^k(x_i)$ denotes k nearest neighbours of x_i. The adjacency matrix A contains only the data points of high closeness that are residing in their k nearest neighbour areas. If two data points x_i and x_j are not in the neighbourhood area of each other, their corresponding value a_{ij} will be set to be zero.

When the data points are projected to the new lower dimensional space, there are two possibilities of the learning process.

Case 1. If two data points x_i, x_j are far in the original space, i.e., $a_{ij} = 0$, the optimizing term in Eq. (2.2) will not bring any effect to the learning process.

Case 2. If two data points x_i, x_j are close in the original space, i.e., $a_{ij} > 0$, their new representations h_i, h_j should share the same neighbourhood. Hence, their distance $\|h_i - h_j\|^2$ should be as small as possible. The optimizing term in Eq. (2.2) will minimize the distance between two representations h_i and h_j when $a_{ij} > 0$ (i.e., x_i and x_j are close in the original data space). Eq. (2.2) can be equivalently written as below,

$$\Leftrightarrow \min\Big(\sum_{i=1}^{n} h_i h_i^{'} \sum_{j=1}^{n} a_{ij} + \sum_{j=1}^{n} h_j h_j^{'} \sum_{i=1}^{n} a_{ij} - 2\sum_{i=1}^{n_h}\sum_{j=1}^{n_l} h_i h_j^{'} a_{ij}\Big) \tag{2.4}$$

$$\Leftrightarrow \min\big(2Tr(H^T D H) - 2Tr(H^T A H)\big) \tag{2.5}$$

$$\Leftrightarrow \min Tr(H^T LH) \tag{2.6}$$

where $Tr(.)$ denotes the Trace of a matrix. D is the diagonal matrix where $d_{ii} = \sum_{j=1}^{n} a_{ij}$ and $L = D - A$ is the Laplacian graph. The minimizing term in Eq. (2.6) is called manifold learning [11, 21]. This manifold learning is combined with NMF framework in clustering to find accurate and meaningful clustering solutions while maintaining the original structure of data. Since NMF factorizes the higher-order data matrix into smaller matrices, it applies many approximations [29]. Incorporating manifold learning into the NMF framework helps approximating the factor matrices that maintain the geometric structure of the original data and produce the meaningful clusters.

For the traditional data, GNMF [21] was the first method to incorporate manifold learning into an NMF framework to learn the low-rank representations that preserve the local geometric structure. Due to maintaining distances between neighbouring points when projecting to lower order, close points in original space are guaranteed to share the same cluster. This helps, resulting in the meaningful cluster solution. The objective function of GNMF on traditional one-view data is defined as below,

$$\min \|X - HW\|_F^2 + Tr(HLH^T), \text{ s.t. } H \geq 0, W \geq 0 \tag{2.7}$$

where L is the Laplacian graph, constructed as in Eq. (2.6).

2.3 NMF-based Clustering Methods on Multi-Aspect Data

The previous section has presented an overview of multi-aspect data characteristics as well as using the NMF framework in learning the meaningful low-rank representation for traditional data. This section will present clustering methods on multi-aspect data based on the NMF framework.

2.3.1 NMF-based Clustering Methods on Multi-view Data

Consider the multi-view dataset $X = \{X_1, X_2, ..., X_{n_v}\}$ with n_v views in total. Let data in vth view be represented as the data matrix $X_v \in \mathbb{R}_+^{n \times m_v}$ where n is the number of samples and m_v is the number of features in vth view. The multi-view clustering task is to group data samples into meaningful clusters by utilising all complementary and compatible views data in the learning process.

The extension of the NMF framework on multi-view data can be handled by applying NMF to learn the low-rank representation of data on each view and seek the consensus representation, i.e., the common latent features for data samples. The consensus representation is then used as the input to a traditional clustering method to achieve the clusters of data respecting data representations on different view features. While the use of NMF on each data view is straightforward, clustering methods have different approaches to learn the accurate consensus matrix.

The most popular and well-known multi-view NMF framework is MultiNMF [56], which is used in many later works [83, 85, 114, 121]. The objective function of MultiNMF [56] can be described as follows:

$$\min \sum_{v=1}^{n_v} \|X_v - H_v W_v\|_F^2 + \sum_{v=1}^{n_v} \|H_v - H_*\|_F^2, \text{ s.t. } H_v \geq 0, H_* \geq 0, W_v \geq 0 \quad (2.8)$$

where $H_v \in \mathbb{R}_+^{n\times r}$ is the new low-rank representation of data corresponding to the basis $W_v \in \mathbb{R}_+^{r\times m_v}$ under the vth view, r denotes the number of the new rank and $H_* \in \mathbb{R}_+^{n\times r}$ is the consensus latent feature matrix of all views. In this objective function, the consensus matrix H_* is effectively learning at the same time as the low-rank representations are learned via the factorizing step.

Another approach is to learn the low-dimensional representations for all views as the first step and then learn the consensus matrix after all view factor matrices have been learned. The objective function of methods [85, 102, 103, 114] adopting this approach is given as bellow,

$$\min \sum_{v=1}^{n_v} \|X_v - H_v W_v\|_F^2, \text{ s.t. } H_v \geq 0, W_v \geq 0 \tag{2.9}$$

The optimizing process in this objective function, similar to the conventional NMF objective function [64], is updating $\{H_v\}_{v=1..n_v}$ and $\{W_v\}_{v=1..n_v}$. This results in generating optimal low-rank matrices H_v and W_v such that $H_v W_v$ is a good approximation of X_v for all $v = 1..n_v$. In the fusion step, the consensus latent feature matrix of all views, denoted as H_*, is linearly combining all individual low-rank representations as in [102],

$$H_* = [H_1 ... H_{n_v}] \tag{2.10}$$

or calculated by taking average as in [103],

$$H_* = \sum_{v=1}^{n_v} H_v / n_v \tag{2.11}$$

The objective function in both Eqs.(5.4) and (2.9) is able to simultaneously learn a low-rank data representation from each data view. In the later step, the consensus data matrix will be learned via compensating the newly learned data representations from all views.

2.3.2 NMF-based Clustering Methods on MTRD Data

The MTRD setting has a small difference in the definition and formulation to the multi-view setting. Consider an MTRD dataset $\mathcal{D} = \{X_1, X_2, ..., X_m\}$ with m object types. Let each object type $X_h = \{x_i\}_{(1 \leq i \leq n_h)}$ be a collection of n_h objects. Let $\{R_{hl} \in \mathbb{R}_+\}$ be a set of inter-type relationship matrices where $R_{hl} = \{r_{ij}\}^{n_h \times n_l}$, r_{ij} denotes the inter-type relationship (e.g., the *tf-idf* weight of a term in a document) between ith object and jth object of X_h and X_l, respectively, $R_{lh}^T = R_{hl}$.

The MTRD clustering task is to simultaneously cluster m object types into clusters by considering all possible relationships. Early NMF-based methods for MTRD make use of the inter-type relationship only [70] and utilize the non-negative tri factorization framework (NMTF) [29] to formulate the objective function as follows,

$$J_1 = \min \sum_{1 \leq h < l \leq m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2, \text{ s.t., } G_h \geq 0, G_l \geq 0 \tag{2.12}$$

where $\|.\|_F$ denotes the Frobenius norm and G_h and G_l are low-rank representations of object types X_h and X_l respectively. S_{hl} is a trade-off matrix that provides an additional degree of freedom to the factorization process [29].

A variation using the Symmetric Non-negative Matrix Tri-Factorization (STNMF) objective function [99] has also been used in some methods [48, 66] as,

$$\min \|R - GSG^T\|_F^2, G \geq 0 \tag{2.13}$$

where the different inter-type relationships of different object types are encoded in

a symmetric matrix R and G is the symmetric low-rank matrix defined below,

$$R = \begin{bmatrix} 0^{n_1 \times n_1} & R_{12}^{n_1 \times n_2} & \cdots & R_{1m}^{n_1 \times n_m} \\ R_{21}^{n_2 \times n_1} & 0 & \cdots & R_{2m}^{n_2 \times n_m} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m1}^{n_m \times n_1} & R_{m2}^{n_m \times n_2} & \cdots & 0^{n_m \times n_m} \end{bmatrix} \tag{2.14}$$

$$G = \begin{bmatrix} G_1^{n_1 \times c} & 0^{n_1 \times c} & \cdots & 0^{n_1 \times c} \\ 0^{n_2 \times c} & G_2^{n_2 \times c} & \cdots & 0^{n_2 \times c} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_m \times c} & 0^{n_m \times c} & \cdots & G_m^{n_m \times c} \end{bmatrix} \tag{2.15}$$

where c is the number of clusters, m is the number of object types (same as the number of views plus 1). The trade-off matrix S in Eq. (2.17) also has the symmetric form similar to R.

As stated before, in order to achieve a high performance in the learning process, the NMF framework should be integrated with the manifold learning for the benefit of learning representation respecting the original geometric structure. The next part will focus on reviewing a wide range of NMF and manifold learning-based clustering methods for multi-view and MTRD data.

2.4 NMF and Manifold Learning-Based Clustering Methods on Multi-Aspect Data

This section will start by analysing the general objective functions designed for multi-view and MTRD using NMF and the manifold learning approach. After that, a detailed review on this approach is presented in three categories, based on how the manifold learning technique is applied in different learning methods for multi-aspect data.

For multi-view data, the NMF objective function to learn the latent features with preserving the local geometric structure on all views is defined as [83, 117],

$$\min \sum_{v=1}^{n_v} \left(\|X_v - H_v W_v\|_F^2 + Tr(H_v L_v H_v^T) \right), \text{ s.t. } H_v \geq 0, W_v \geq 0 \tag{2.16}$$

where $L_v = D_v - A_v$ is the Laplacian graph and A_v is the adjacency matrix built on the vth data view. The second term in this objective function helps the learning process to return the low-rank latent feature matrices $\{H_v\}$ that are smooth with the intrinsic geometric structure of data view vth and thus can be more meaningful as compared to latent feature matrices $\{H_v\}$ obtained from the multi-view NMF objective function in Eq. (2.9).

On MTRD data, the NMF objective function to learn the latent features with the local geometric structure preserved on all types is defined as [99, 100],

$$\min \|R - GSG^T\|_F^2 + \lambda Tr(G^T L G), \text{ s.t. } G \geq 0 \tag{2.17}$$

where the Laplacian L is symmetrically defined as,

$$L = \begin{bmatrix} L_1^{n_1 \times n_1} & 0^{n_1 \times n_2} & \cdots & 0^{n_1 \times n_m} \\ 0^{n_2 \times n_1} & L_2^{n_2 \times n_2} & \cdots & 0^{n_m \times n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_m \times n_1} & 0^{n_m \times n_2} & \cdots & L_k^{n_m \times n_m} \end{bmatrix} \tag{2.18}$$

L_1 is the Laplacian defined on sample object type, L_2 is the Laplacian defined on feature object type 1 and so on. The symmetric inter-type relationship R and the symmetric factor matrices G are defined as in Eqs. (2.14) and (2.15), respectively. The objective function in Eq. (2.17) utilizes NMF to simultaneously cluster many object types and incorporated Laplacian matrices to learn manifolds on sample object types and all feature object types.

Applying manifold learning on NMF-based clustering methods for multi-aspect

data is not a trivial extension of traditional data. Since data on different aspects may sample on different structures, there are three main family of methods according to the way they deal with this challenge.

(1) A natural extension of manifold learning on multi-aspect data is learning the intrinsic manifold of data on each aspect, i.e., view or object type, and ensuring the low-rank data in the new mapped space is smooth with the intrinsic manifold on the corresponding aspect. This first family of methods focuses on learning and preserving the manifold on each aspect to learn different aspect factor matrices meaningfully.

(2) Similar to using the manifold learning on traditional data, the most important thing is to learn the accurate manifold embedded on each aspect and constrain the low-rank representation on each aspect to lie on the corresponding intrinsic manifold. This will help to achieve the accurate cluster structures from each aspect by learning the correct factor matrices. This second family of methods relies on learning the accurate manifold on each aspect to learn the meaningful factor matrices.

(3) Some state-of-the-art methods regard that the intrinsic manifold in the dataset is embedded in a convex hull of all the manifolds of all views [66, 117]. The convex hull of all manifolds from all views can be the minimal convex set containing all manifolds or a convex combinations of all manifolds [38]. After learning the consensus manifold, it is a vital problem to learn the low-order representations to ensure the smoothness with the intrinsic manifold learned from all views. Therefore, this emerging family of methods focuses on learning the consensus manifold embedded from multiple manifolds from all views. The methods rely on the learned consensus manifold to find the consensus coefficient matrix.

Next, we provide a comprehensive review of the three classes of NMF and manifold learning-based clustering methods on multi-aspect data.

2.4.1 Learning the Manifold on Each Aspect

An easy extension to transform the application of manifold learning from a single aspect to multi-aspect is learning the manifold of data on each aspect and preserving the manifold when learning the low-dimensional data on each aspect.

In multi-view data, inspired by the idea that different views data should contain similar embeddings structures, the authors in [114] proposed the objective function that attempts to learn the low-rank representations from all views such as these representations is as similar as possible. At the end of the optimizing process, it is believed that each coefficient matrix will become the consensus coefficient matrix and will embed the cluster structure. In order to ensure the meaningfulness of the consensus matrix, the graph regularization is incorporated into the objective function to help learn the low-rank representations respecting their corresponding manifolds. The objective function is defined as,

$$\begin{aligned} \min \big(& \sum_{v=1}^{n_v} \|X_v - H_v W_v\|_F^2 + \sum_{v=1}^{n_v} \sum_{s=1}^{n_v} \|H_v - H_s\|_F^2 \\ & + \alpha \sum_{v=1}^{n_v} \|(H_v)^T - I\|_F^2 + \lambda \sum_{v=1}^{n_v} Tr(H_v L_v H_v^T) \big), \\ & \text{s.t. } W \geq 0, H \geq 0 \end{aligned} \tag{2.19}$$

The first term in the objective function requires H_v to be the best low-rank representation learned from data view X_v. The second term requires that coefficient matrix H_v must be similar to other coefficient matrices of other views. The third term $R = \min \sum_{v=1}^{n_v} \|(H_v)^T - I\|_F^2$ is added to emphasize the orthogonality constraint on factor matrix H_v. This constraint enables a return to the unique and interpretable matrix [29]. The last term is to guarantee the manifold assumption of every data point on the low dimensional space, i.e., the low-rank data points should lie on the embedded manifold of the corresponding view.

It can be noted that the low-rank representation H_v from each view is restricted by using many constraints in objective function Eq. (2.19). This helps returning in the good low-order representations on the well-defined datasets, yet it can cause bad results on some datasets because of too many constraints.

In the context of MTRD, Dual Regularized Co-Clustering (DRCC) [43] is designed for co-clustering data, i.e., simultaneously clustering samples and features in a dataset where data samples are represented by one type feature. Co-clustering, also known as clustering on bi-type data, is based on the duality between samples and feature and is a special case of MTRD (i.e., two-type data) clustering. DRCC is the first method designed for learning manifolds on both data samples and data features through the co-clustering problem [28]. It investigates the structures of features and the clustering of features boosts the clustering process of samples. When the more accurate feature clusters are learned, the more accurate sample clusters will be achieved. Therefore, simultaneously seeking the low-rank representations of samples and features as well as simultaneously maintaining the geometric structures of low dimensional spaces on both samples and features have shown to be the most effective learning process for the co-clustering problem [28].

In DRCC, two graphs, one for data samples and another for data features, were constructed to model the local geometric structure information. This graph regularization is incorporated in the NMF objective function to introduce the smoothness of both data points and data features with their embedded manifolds. The objective function was defined as below,

$$\begin{aligned} &\min(\|X - HSF^T\|_F^2 + \lambda Tr(H^T L_H H)) + \mu Tr(F^T L_F F)) \\ &\text{s.t. } H \geq 0, F \geq 0 \end{aligned} \tag{2.20}$$

The l_2-norm constraint is assigned on rows of the learned low-ranked matrices H and F to ensure the objective function is lower bounded and the trade-off matrix S is relaxed to take any signs.

As a formal extension of DRCC, STNMF [99] proposed learning manifolds for all object types in a MTRD dataset. STNMF was the first method utilising NMF to simultaneously cluster many object types. STNMF proposed a novel symmetric framework for the MTRD clustering problem and incorporated Laplacian matrices to learn manifolds on sample object types and all feature object types. The objective function is defined as,

$$\min \|R - GSG^T\|_F^2 + \lambda Tr(G^T LG), \text{ s.t. } G \geq 0 \tag{2.21}$$

where the Laplacian L is symmetrically defined as,

$$L = \begin{bmatrix} L_1^{n_1 \times n_1} & 0^{n_1 \times n_2} & \cdots & 0^{n_1 \times n_m} \\ 0^{n_2 \times n_1} & L_2^{n_2 \times n_2} & \cdots & 0^{n_m \times n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_m \times n_1} & 0^{n_m \times n_2} & \cdots & L_k^{n_m \times n_m} \end{bmatrix} \tag{2.22}$$

L_1 is the Laplacian defined on the sample object type, L_2 is the Laplacian defined on the feature object type 1 and so on. The symmetric inter-type relationship R and the symmetric factor matrices G are defined as in Eqs. (2.14) and (2.15), respectively.

2.4.2 Learning the Accurate Manifold on Each Aspect

Manifold learning aims at preserving the local or global geometric structures of data [97]. Since locality preserving shows a tendency to group similar data, most manifold learning methods incorporated in clustering aim at preserving the local structure of the data. These manifold learning methods rely on building a kNN graph [11], a powerful technique used widely in machine learning and data mining. The kNN graph models the neighbourhood information for each data point and encodes the closeness between each data point to all its neighbours. This closeness information will be embedded in the learning process to make sure the projection preserves the

local geometric structure of data. The calculation of closeness of data objects can be varied depending on different data types. The pairwise similarity, calculated based on different similarity weighted schemes such as Euclidean distance, Cosine similarity or Gaussian kernel [11, 21], is most common.

Relational multi-manifold co-clustering (RMC) [66] was proposed, in an effort to learn the accurate intrinsic manifold for MTRD data. RMC learns the optimal manifold from a combination of many predefined manifolds. The predefined manifolds are built as initial guesses of graph Laplacian. Different guesses can be constructed by different kinds of weighted schemes or different values for the neighbourhood size parameter. In RMC, the intrinsic manifold is calculated by the following equation,

$$L = \sum_{i=1}^{q} \mu_i \widetilde{L}_i, \text{ s.t.} \sum_{i=1}^{q} \mu_i = 1, \mu_i \geq 0 \tag{2.23}$$

where $\widetilde{L}_i$ denotes a candidate manifold i and q is the number of candidate manifolds. Each candidate manifold is a combination of different manifolds of different object types expressed as the symmetric form equation,

$$\widetilde{L}_i = \begin{bmatrix} L_1 & 0 & 0 \\ 0 & L_2 & 0 \\ \dots & \dots & \dots \\ 0 & 0 & L_m \end{bmatrix} \tag{2.24}$$

L_1 is the Laplacian of samples data and different L_i are different manifolds of different feature object types. The candidate Laplacian graph $\widetilde{L}_i$ will be embedded in the symmetric framework with the following objective function,

$$J = min\|R - GSG^T\|_F^2 + \alpha Tr\big(G^T(\sum_{i=1}^{q} \mu_i \widetilde{L}_i)G\big) + \beta\|\mu\|^2 \tag{2.25}$$

R and G are non-negative and formulated as in Eqs. (2.14) and (2.15) respectively.

l_2-norm on μ is for the even distribution of the parameter for all manifolds and it prevents the parameter overfitting to one manifold. Since RMC considers several possible manifolds, the learned consensus manifold is believed to be closest to the intrinsic manifold of original data. It learns many manifolds; however they are of the same type, i.e., based on kNN graph. Consequently, the learned manifold in RMC is less diverse as well as it incurs an extra computational cost to calculate the ensemble manifold. In addition, the parameter for the neighbourhood size k for building the kNN graph can not be known apriori. This makes the value k a user defined parameter, which needs to be optimised. A big value of k may lead to include non-useful neighbours in the kNN graph that can lead to learning an inaccurate manifold, whereas a small value of k may not include all useful neighbour points that will lead to learning an incomplete manifold. Knowing an optimal value for k helps to choose all useful neighbours for constructing the kNN graph. It helps to learn a more meaningful affinity matrix that is required to learn the accurate manifold and achieve a meaningful clustering solution.

To ensure learning a more diverse manifold or ensure all useful neighbour information is included in constructing the affinity matrix, RHCHME [48] was proposed to build the manifold by combining Euclidean distance learning and subspace learning. The complete manifold is learned as,

$$L = \alpha L^E + L^S \tag{2.26}$$

where $L^S = D - A^S$ with D is the diagonal matrix where $(d)_{ii} = \sum_j (a^S)_{ij}$. A^S is a similarity matrix learned from considering data points lying on subspaces, i.e., two data points will be considered as neighbours if they belong to the same subspace, despite the distance between them [39]. $L^E = D - A^E$ with $(d)_{ii} = \sum_j (a^E)_{ij}$. A^E is the affinity matrix derived from constructing kNN graph as in Eq. (2.3).

The Laplacian matrix constructed in RHCHME can learn a more comprehensive

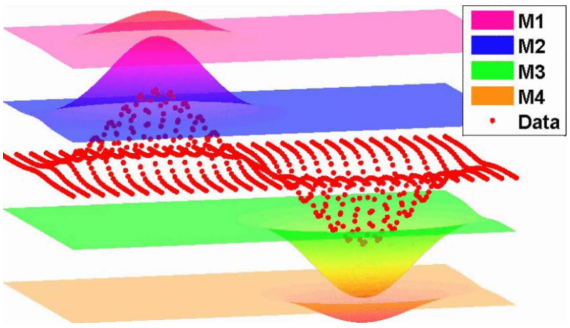


Figure 2.4: An example of multiple manifolds and the convex hull of manifolds [38]. Suppose M1, M2, M3, M4 are different manifolds learned from different views, the consensus manifold of the multi-view data can be similar to how the original data is sampling, i.e., the data manifold (the red dot shape), which is a convex combination of all manifolds.

intrinsic manifold for data when projecting. The objective function of RHCHME [99] is based on the STNMF framework and is similar to RMC [66]. However, the Laplacian graph is not an ensemble of many candidate manifolds but is built by considering the data lying on the manifold as well as belonging to subspace. The more comprehensive manifold learned in RHCHME helps to provide more meaningful clusters evidenced by more accurate clustering results. However, similar to other clustering methods relying on manifold learning, RHCHME learns the manifold by considering the local geometric structure only. The local geometric-based learning manifold aims at preserving the close distances between neighbour points and/or ensuring membership to subspace of data points.

These methods avoid considering distance information of data points that are not in the neighborhood, due to the computational cost and usefulness of the local distance in clustering rather than using the global distance. However, it has been shown that preserving all distance information both for close and far points will help to learn more meaningful representations [97]. A more comprehensive manifold may be needed to help the learning process to achieve the more meaningful clusters.

2.4.3 Learning the Intrinsic Consensus Manifold in Multi-View Data

Data from all views sampled from different manifolds should lie on the intrinsic manifold that is the convex hull of the manifold of all views [117]. This convex hull is the minimal convex set containing all manifolds or a convex combination of all manifolds [38]. An example of the convex combinations of all manifolds is depicted as in Figure 2.4. When using manifold learning on multi-view data, in addition to the difficulty of learning the accurate manifold on each view, multi-view learning should also deal with how to learn the accurate consensus manifold from multiple views and then learn the meaningful consensus coefficient matrix respecting the consensus manifold. It has been evidenced in the literature that it is not a trivial task to effectively employ manifold learning for NMF based multi-view data for effective understanding. To the best of our knowledge, MMNMF [117] is the only method learning the consensus manifold and using it in the NMF framework effectively. The method proposes two objective functions to learn the accurate coefficient matrix relying on the consensus manifold. The consensus manifold is learned beforehand by linearly combining the manifolds learned from all views as,

$$L_* = \sum_{v=1}^{n_v} \lambda_v L_v \tag{2.27}$$

The first objective function (MMNMF1) is defined as,

$$J = \sum_{v=1}^{n_v} D_{KL}(X_v||H_v W_v) + \lambda Tr(H_*^T L_* H_*) \tag{2.28}$$

where $D_{KL}(.)$ is the Kullback-Leiber Divergence. The consensus coefficient matrix H_* is constructed by linearly combining all views low-rank matrices as $H_* = \sum_{v=1}^{n_v} \alpha_v H_v$. In this objective function, different coefficient matrices are learned in the new manner to ensure the local geometric structure is embedded on each view.

Therefore, it is deemed that the consensus matrix will be ensured to preserve the diverse geometric structures for the newly mapped space of all views and will return a diverse consensus matrix. However, it can also be considered as a disadvantage of the objective function since it will balance the importance of different views and fail to learn the accurate consensus matrix if some views in the dataset are more important than others.

In the second objective function (MMNMF2), the common coefficient matrix is learned simultaneously as the low-rank matrices from all views are learned. The objective function [117] is expressed as,

$$J = \sum_{v=1}^{n_v} D_{KL}(X_v||H_v W_v) + \alpha_v D(H_* - H_v) + \lambda Tr(H_*^T L_* H_*) \tag{2.29}$$

It can be noted that different factor matrices during the learning process are constrained to be as close to the consensus matrix as possible. At the same time, it is ensured that the data points belonging to the consensus low dimension space should lie on the convex hull of the multi-manifold. The consensus matrix from this objective function can better reflect the latent clusters shared from all views' data. MMNMF2 can be considered as the well-defined framework of combining NMF and manifold learning since it is able to return the natural consensus shared from all views that considers the consensus manifold embedded.

MMNMF is believed to be the most effective method since it considers all cases of applying the manifold on multi-view data. However in this method, since the consensus manifold is linearly combining all manifolds from all views, it may fail to reflect the natural consensus manifold. Similar to learning the consensus coefficient matrix, the consensus manifold should also be learned naturally during the learning process.

The MMNMF method is designed for multi-view data. There are no methods exploiting the consensus manifold designed for MTRD or partial multi-view data.

Furthermore, the NMF-based multi-view method utilising the consensus manifold should also be combined with the techniques to learn the accurate manifold on each view, in order to achieve the best solution.

2.5 Other Multi-Aspect Data Clustering Approaches

2.5.1 Feature Concatenation based Multi-Aspect Data Clustering

A naive approach for multi-view data clustering is concatenating all data views into a single view and then applying a traditional clustering method on this view. Since the concatenation of all views includes all data, it is assumed that clustering will utilize the structures of the data of all views. However, the concatenation not only disregards the interrelatedness between data but also causes poor performance as the single concatenating data will be very sparse and high dimensional. As shown by researchers [14, 70], the approach of using concatenation for all views data fails to improve the clustering performance.

2.5.2 Spectral Clustering Approach

Spectral clustering algorithms [28, 77] are one of the effective clustering methods that are based on the established and well-studied spectral graph theory [32]. It is based on graph partitioning and aims to find the best cuts of a data graph. Proposed in previous methods [9, 77], it has been applied effectively on traditional single-view data. A variation of this method can be considered for two-type data where it builds the original data as a bipartite graph and finds optimal partitions by utilising the association relationship between sample clusters and feature clusters [28]. Consistent Bipartite Graph Co-Partitioning (CBCC) [33], a spectral clustering

on two-view data, is designed to be applied on interrelated multi-feature data. It aims to include the star-structured inter-relations with the simplest star-structured case when the sample objects are at the centre and are represented via two different feature objects. By treating the data tripartite graph as two bipartite graphs and searching the clusters for a central type by a fusion process from the two co-clustering problems, it has shown the effectiveness over its counterparts of single-view data. Since being the first work extended for high-order co-clustering, it was not scalable to large-scale data sets.

The majority of spectral-based methods work in the following manner. The problem is described in such a way that the spectral theory can be applied, while the objective function is designed so that it can be relaxed to have a variant of Rayleigh quotient to find the best cut of the similarity graph via the k largest eigenvectors. In the context of multi-view learning, methods work under the assumption that data samples represented through many different feature types should have the same cluster structures. Since the spectral clustering aims partition the similarity graph to different parts, the multi-view spectral clustering aims to learn the consensus partitions among the similarity graphs of all views.

The majority of multi-view spectral clustering methods use the famous well-defined spectral clustering framework [77] as the basis and focus of learning the optimal consensus graph cut. Steps in [77] to learn k cluster labels of a data matrix X can be summarized as follows,

- Step 1: Build the similarity graphs based on the data input
- Step 2: Build the normalized Laplacian matrix and its k largest eigenvectors from the affinity matrix using the previous step
- Step 3: Apply k-means on the newly built space formed by using k largest eigenvectors as columns and obtain the clustering indicator matrix

In general, the extension to multi-view should follow these steps. However, since

the aim of spectral based multi-view clustering is finding the similar partitions of all graph views, the consensus partitions can be obtained after learning the partitions from all views such as [62], or learning the consensus Laplacian matrix as in [59, 68].

The co-regularized spectral method [62] can be considered as the typical spectral based multi-view clustering method. Spectral clustering is first applied on each view data. A consensus eigenvector matrix that embeds eigenvector structures of all data views is then obtained. To find the consensus eigenvector matrix, the co-regularization idea from supervised learning is used that measures the disagreement between clustering of different views. The objective function designed for multi-view clustering is solved by spectral clustering as follows,

$$\begin{aligned} &\max \sum_{v=1}^{m} Tr(U_v^T L_v U_v) + \sum_{v} \lambda Tr(U_v U_v^T U_* U_*^T), \\ &\text{s.t. } U_v^T U_v = I, U_*^T U_* = I \end{aligned} \tag{2.30}$$

The multi-view spectral clustering aiming to learn the consensus Laplacian graph will require learning of Laplacian graphs of all views and finding the consensus among all views. This approach such as MVSC [68] is known to be time consuming. The spectral objective function of MVSC is given as follows,

$$\min \sum_{v=1}^{n_v} (\alpha_v) Tr(H^T L_v H), \text{ s.t., } G^T G = I, \sum_{v=1}^{n_v} \alpha_v = 1, \alpha_v \geq 0 \tag{2.31}$$

The complementary information of all views is embedded through L_v. The consensus matrix H is learned simultaneously during the learning process through all views. MVSC deals with the scalability issue by using a set of salient points connecting with data samples as a bipartite graph and simultaneously clustering salient points and sample points.

Recently, MVSC-CEV (Multi-view Spectral Clustering by Common Eigenvectors) [59] has been proposed based on the assumption that different views data

should share similar k largest eigenvectors. In the second step of seeking the eigenvectors of all similarity graph views, MVSC-CEV makes use of S-CPC (Stepwise common principle components) [98] to find the common eigenvectors of all Laplacian matrices and use these common eigenvectors for clustering, as in step 3 of the traditional spectral clustering approach.

The Spectral Relational Clustering (SRC) [70] method is developed by combining the concepts of spectral clustering and NMF. It represents the multi-type relational data by collectively factorizing data matrices and solving the MTRD clustering by using spectral graph theory. So far, SRC is the most well-defined spectral framework for MTRD but it fails to involve intra-type relationships.

Though spectral clustering is well-defined since it relies on spectral graph theory and has been proved to be an effective approach, yet the spectral clustering methods inherit the limitation of graph theory which is failing to work on large-scale data sets.

2.5.3 Subspace Learning

In the last few years, subspace learning has gained much attention in multi-view learning due to its ability to learn the latent subspace of data points. This approach assumes that data points could be drawn from multiple subspaces. Since the multi-view data is very sparse and high-dimensional, dimension-reduction based subspace learning can be an effective approach for the multi-view data to deal with the "curse of dimensionality". Examples of state-of-the-art multi-view subspace clustering methods are Multi-view subspace clustering (MVSC) [36], Latent Multi-view subspace clustering [113] and MLRSSC [19]. The process includes two main steps: seeking the latent subspaces of data and then applying a traditional clustering method on the newly learned subspaces.

Since data samples of different views may distribute in different subspaces, the vital step would be to find the unified subspace. Different approaches have been

used to find the unified subspace of data. MVSC [36] performs subspace learning on each view and constrains the latent subspaces of each view to lie on the same block structures of data by making use of the cluster indicator matrix in the constraint. DiMSC (Diversity-induced Multi-view Subspace Clustering) in [22] is a combination of subspace learning and manifold learning on all views, where the diversity constraint is added to boost the diversity or complementary information between the views in order to achieve meaningful representations. Most multi-view methods are based on an additional regularization as an effort to accurately learn the consensus subspace. Different from the above methods that are based on the assumption of the complementary information of all views, the subspace learning-based multi-view method in [105] relies on exploiting the consensus correlation of data from all views, i.e., the data points should have the same levels of relatedness with regard to subspaces across all views. This assumption brings the meaningful unified representation since the data points in the same subspace have encoded the high correlations and are therefore producing more accurate clusters.

As mentioned in early works [4, 79], subspace learning can be considered as NMF's counterpart for the dimensionality reduction task, however, it is important to note that, while NMF framework works as a feature transforming technique, subspace learning performs feature selection. For the high relatedness data (e.g., text data) that the thesis aims at, NMF is selected in order to retain the relation from original data when projecting to lower-order space.

2.6 Research Gap and Summary

This chapter has reviewed the literature relevant to multi-view clustering methods focusing on NMF and manifold learning. The following research gaps have been identified.

2.6.1 Developing a Useful Characterization of the NMF and Manifold Learning-Based Methods for Multi-Aspect Data

Researchers have proposed a number of methods to effectively understand multi-aspect high dimensional data ranging from focusing on learning the accurate manifold on each object types in MTRD [48, 66] to learning the correct consensus manifold from multiple views after having a manifold learned on each view [120] in multi-view data. This is an emerging research field that needs attention and that is required to be presented in a systematic fashion. However, there has been no comprehensive discussion of this specific field of multi-aspect clustering and associated methods to provide a general view of how to understand a multi-aspect problem via applying NMF and a learning manifold, or to discuss the challenging problems that can be addressed in future research.

2.6.2 Learning the Accurate Manifold

Combining NMF and manifold has gained success on both traditional and multi-view data over the last decades. However, using manifold learning in multi-aspect is challenging. An incorrect manifold may cause a bad effect on the learning process as well as it not being a trivial task to learn the accurate manifold for both homogeneous and heterogeneous data. The challenges of using manifold learning can range from choosing a precise neighbourhood size to learn the local geometric structure, to learning the manifold that can embed both local and global information. Specific to multi-view applications, the challenge includes learning the low-rank representations of each view, respecting the manifold of the corresponding view as well as respecting the intrinsic consensus manifold for multi-view data. To the best of our knowledge, most existing manifold-based methods have limited ability to dealing with the above-

mentioned difficulties.

2.6.3 Learning Diverse Relationships

As noted earlier, the multi-view/MTRD data exhibits various types of relationships including intra- and inter-type relationships, intra- and inter-view relationships and cluster association relationships. While inter-type, intra-type and intra-view relationships have been extensively exploited in MTRD/multi-view data, the inter-view relationship is often neglected. One may need a new formulation for using an inter-view relationship in the MTRD-based framework along with all intra-view relationships. The association relationship has been included in the spectral-based method; however, it is needed to be exploited effectively in low-rank-based approaches such as in subspace and NMF methods. Considering association relationships in the learning process will not only help to learn a general view of how different clusters are related to each other in heterogeneous data, but it can also help the factor matrices to be optimized faster to the optimal solutions.

2.6.4 Learning the Consensus Matrix in Multi-view Data

In the low-rank-based multi-view learning, the consensus matrix of the latent feature is learned independently from the low-rank matrices generated for all views. This approach of existing methods has shortcomings such as: 1) the step of learning low rank representations from all views is normally conducted independently for each view. Consequently, it has a high chance that the salient information relating to the relationships between views will be lost; and 2) the step of learning the consensus from all low-rank representations may disfavour the levels of contribution of each data view and may lead to a non-realistic outcome. A new design for the low-rank learning framework, where the consensus matrix is learned naturally and simultaneously during the learning process together with the low-rank representations of all

views, is needed in order to achieve the meaningful cluster structures.

2.6.5 Learning both Compatible and Complementary Information for Multi-view Data

The intent of multi-view data is the inclusion of not only compatible but also complementary data from all views. Most existing methods favour either the compatible or complementary only, the reason may lie in the complex nature of unsupervised learning or the limitation of the framework used. However, failing to learn both properties may affect the production of an accurate clustering outcome. A new designed framework for multi-view data that can allow the learning process to learn and maintain the crucial compatible and complementary adequately is more than necessary.

2.6.6 Learning the Consensus Manifold for Multi-view Data

Multi-view data may sample on different manifolds, however, due to the compatible and complementary principles, there should exist a consensus manifold that lies on the convex hull of all manifolds corresponding to all views. Learning the consensus manifold for multi-view data is a non-trivial task. Learning a correct consensus manifold and ensuring the consensus low-rank representation to respect this manifold, is beneficial to the multi-view learning process by returning a more accurate and more meaningful representation. Unfortunately, there is a lack of methods proposed regarding learning the consensus manifold.

2.6.7 Missing Applications of Multi-Aspect Data Learning to Other Related Real-world Problems

The multi-aspect data approach has been shown to bring rich information to unsupervised learning. Developed multi-aspect learning methods can be applied in many other related application domains such as community discovery or collaboration filtering, however, not many works exist that deploy multi-aspect data learning methods to other real-world problems.

Chapter 3

Learning Association Relationship and Accurate Geometric Structures for Multi-type Relational Data and Its Application to Community Discovery

This chapter introduces the proposed method named ARASP (Learning Association Relationship and Accurate Geometric Structures for Multi-type Relational Data) and an application of ARASP on the problem of community discovery. These two contributions will comprise this chapter, in the form of two papers.

Paper 2. **Khanh Luong** and Richi Nayak (2018), *Learning association relationship and accurate geometric structures for multi-type relational data.* In 34th International Conference on Data Engineering (ICDE 2018), Paris, France.

Paper 3. T.M.G, Tennakoon, **Khanh Luong**, Wathsala Mohotti, Sharma Chakravarthy, and Richi Nayak, *Multi-type Relational Data Clustering for Community Detection by Exploiting Content and Structure Information in Social Networks.* The 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019) (In Press).

The first part will present the proposed method, ARASP, which focuses on dealing with the challenge of learning the accurate geometric structure to aid in the MTRD learning based on the NMF framework. In this paper, an important relationship type, the association relationship, which has been ignored or inadequately considered in other methods, has been effectively included into the NMF framework. To deal with learning the accurate geometric structure for MTRD, a novel fashion of learning the affinity matrix for each object type is proposed where both the close and far distance information have been incorporated. It is widely accepted that, only close distances reflect the true non-linear structure, however, far distances, learned using farthest neighbour graph, will help ensuring far points be kept far apart and not be mistreated as relevant points when constructing affinity matrix for manifold learning. This will help in learning a more meaningful representation for MTRD and will be the necessity for a meaningful clustering outcome. To embed the association relationship into the learning process, the famous spectral based constraint named, Normalized Cut-Type, is applied in the objective function to force the objective function to learn and respect the association relationship that inherits between clusters of different object types. This would help the clustering algorithm to learn a general view of how closely the objects from different clusters of different object types are interacting and would lead to a worthwhile clustering solution.

The second part of the chapter focuses on the application of the ARASP method on the problem of Community Discovery, one of the most important and challenging problems of social network learning, for the Twitter dataset, a popular microblogging service. Taking the advantage of clustering, some works recently deemed that

clustering algorithms can be potentially effective for discovering user communities. This paper will approach the detecting community problem under the MTRD context utilising the fact that data is existing under many types of relationships in the Twitter dataset, thus can be the premise for purposeful community learning. The success of this work on Community Discovery not only brings a new look of utilising MTRD Clustering for meaningful community discovering, but also brings a great potential of the multi-aspect learning for other real-world problems.

Next, the chapter will present two papers. Since this is a thesis by publication, each paper will be presented in its original form. Due to their different formats, there will be some minor format differences. However, these do not alter the content of the original papers.

Paper 2. Khanh Luong and Richi Nayak (2018), *Learning association relationship and accurate geometric structures for multi-type relational data.* In 34th International Conference on Data Engineering (ICDE 2018), 16-19 April 2018, Paris, France.

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors:

Khanh Luong (PhD Candidate): Conceived the idea, designed and conducted experiments, analysed data, wrote the paper and addressed reviewers comments (for published papers) to improve the quality of paper.

Signature: ***Date:*** 24-06-2019

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Signature: ***Date:*** 24-06-2019

ABSTRACT. Non-negative Matrix Factorization (NMF) methods have been effectively used for clustering high dimensional data. Manifold learning is combined with the NMF framework to ensure the projected lower dimensional representations preserve the local geometric structure of data. In this paper, considering the context of multi-type relational data clustering, we develop a new formulation of manifold learning to be embedded in the factorization process such that the new low-dimensional space can maintain both local and global structures of original data. We also propose to include the interactions between clusters of different data types by enforcing a Normalize Cut-type constraint that leads to a comprehensive NMF-based framework. A theoretical analysis and extensive experiments are provided to validate the effectiveness of the proposed work.

KEY TERMS. Non-negative matrix factorization; multi-type relational data; manifold learning; association relationship; k nearest neighbour graph; p farthest neighbour graph;

3.1 Introduction

Multi-type relational data (MTRD), representing objects that exhibit multi types of relationships has received much attention in recent years for knowledge representation and discovery. MTRD can embed rich information, e.g., a bibliographic data includes objects such as conferences, authors and key words; and a web search data includes three kinds of objects namely web pages, search queries and web users. Objects in MTRD can have many kinds of relationships, e.g., browsing relationship type between web pages and users, searching relationship type between search queries and users and the intra-type relationship between users or between web pages. Clustering on MTRD has shown to find more meaningful clusters as compared to clustering on "flat" data that can only capture data and their features [31]. Figure 3.1 illustrates an example of a dataset with three different types of objects:

documents, terms and concepts. Simultaneously clustering documents, terms and concepts will result in more meaningful document clusters.

The MTRD clustering can be defined as a process of grouping similar objects by simultaneously considering multi-type features. The last decade has witnessed the emergence of MTRD clustering methods [48, 57, 99, 116]. Non-negative matrix factorization (NMF) based MTRD clustering methods have been found effective and flexible for analysing high-dimensional data [63, 64]. These methods aim at finding intrinsic lower-order dimensional space and searching clusters in this new space. These methods focus on finding the best approximation of the original data and may fail to consider the sampling on manifold of data [21], i.e., they ignore how the geometric data structure will change after mapping from high to lower dimensional space. Several NMF-based clustering methods therefore rely on manifold learning to obtain the part-based representations of data that ensures smoothness with the intrinsic manifold. The manifold learning process embedded in these methods attempts to preserve the local structure of data by maintaining the distances between points within the neighbourhood area utilising a k nearest neighbour (kNN) graph [11]. The kNN graph, encoding close distances between points in the neighbourhood area is then incorporated as an additional regularized term to be followed in NMF when projecting data to lower order embedded space.

These existing methods ignore preserving distances between points that do not share the same neighbourhood. Consequently, the learning process results in an inaccurate and incomplete manifold for original data. The computational cost to compute and preserve distances between every pair of points is very high and the local geometric is seen as more useful to clustering than global [10]. However, it has been shown that finding an accurate and complete manifold structure of data will bring a more faithful representation [97]. We propose to construct the affinity matrix in a novel fashion such that it completely encodes all distance information from a point to its nearest neighbours and to all its farthest points. The newly

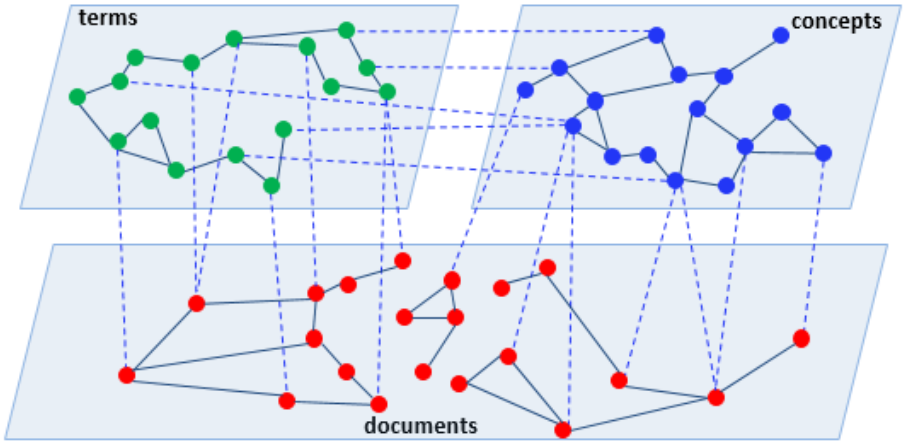


Figure 3.1: An example of MTRD dataset with three object types: documents, terms and concepts. The intra-type relationships are represented as solid lines and inter-type relationships are represented as dotted lines.

constructed affinity matrix will then be formulated as a regularization term to ensure maintaining these distances during the manifold learning process. Since the manifold learning is based on the constructed affinity matrix, the learnt manifold is believed to be more accurate and more complete (considering both close and far points). To our best knowledge, this will be the first work to exploit the accurate data structure by considering farthest points in embedding lower space for MTRD in the NMF framework. This is our first contribution. A major challenge in designing a MTRD clustering method is how to effectively exploit all possible intrinsic relationships that exist in MTRD to obtain more meaningful and accurate clusters [99]. These relationships can be inter-type relationships between objects of different types, intra-type relationships between objects of same type and association relationships between clusters of different object types [70, 99].

While most clustering methods rely on inter and intra relationships as main source of information to cluster objects, the association relationships between clusters are often ignored or considered in an inadequate manner. To our best knowledge, none of the existing methods explore this relationship effectively for MTRD clustering. By including the information about relationships between clusters of different object types would allow the clustering algorithm to learn a general view of how

closely the objects from different clusters of different object types are interacting. The second novel contribution of this work will focus on learning the association relationships among clusters of different object types simultaneously with lower order embedding. This will be done through enforcing non-negative and Normalize Cut-type constraints [93] on factor matrices, leading to the meaningful lower order dimensional space representation.

The proposed MTRD clustering method, named as Association Relationship and Accurate Structure Preserving (ARASP), will be able to provide meaningful and unique cluster representation that includes learning association relationship between clusters as well as ensures smoothness with the intrinsic manifold. The proposed MTRD method results in a well-defined NMF framework that can simply be reduced to apply on traditional data or co-clustering. We also provide the related update rules that guarantee convergence. An extensive empirical analysis on diverse datasets and with several benchmarking methods is presented to validate the effectiveness of the proposed method.

3.2 Related works

NMF is a matrix factorization method focusing on finding an embedded lower dimensional space that can best approximate the original space [63, 64]. NMF-based clustering methods that search for clusters in this new space have shown effectiveness especially on high-dimensional data [21, 29, 42, 43, 81, 109]. Since NMF methods focus on achieving the best approximation of original data, they can fail to preserve the geometric structure of data [21, 34], necessary for finding accurate and meaningful clusters. GNMF [21] was proposed to combine manifold learning with NMF for maintaining the local geometric structure of original data. Inspired by the effectiveness of GNMF on traditional data, DRCC [43] and IGNMF [42] are designed as co-clustering methods incorporating embedded manifold learning for both

data samples and features to cluster rows and columns simultaneously. IGNMF [42] imposes the Normalized Cut-type constraint [93] to make sure the optimization problem is well-defined and no longer suffers from trivial solutions. STNMF [99] is a prominent MTRD NMF-based clustering method that encodes local structure of data on each data type. The method first constructs different affinity matrices for all object types and then combines these matrices into building an affinity graph to be included in the symmetric NMF, for effectively clustering multiple object types simultaneously. STNMF achieved significant improvement compared to traditional clustering method GNMF showing the important role of simultaneously clustering many object types. Yet it is not a trivial task to find exactly embedding manifolds for MTRD, authors in [66] attempted to learn optimal intrinsic manifolds of both sample and feature from some pre-defined candidate manifolds. Some further extensions in [48] uncover the low-dimensional embedded space by considering data samples on manifolds as well as their existences in subspaces, and in [121] by learning embedded multi-manifold of multi-view data spaces. These works have proved the importance and necessity of involving learning manifold into clustering. To the best of our knowledge, there exists no NMF-based MTRD clustering method that saves the accurate structure of data via preserving both close and far distances between data points and improves the clustering performance.

On traditional data, taking the advantage of combining NMF and manifold learning to ensure the intrinsic shape of data when looking for the low-rank representations is preserved, SPNMF [69] discovers the intrinsic geometric shape by discriminating structure of input data through constructing two matrices, i.e., affinity matrix and repulsive matrix encoding high and low relatedness respectively. Unfortunately, the construction of repulsive matrix relies on learnt affinity matrix [24] which requires re-calculating distances and thus lead to re-computing the Laplacian matrix after each iteration. This results in high computational cost and makes the method limited to small size data only. In this work, rather than neglecting far distance

information or constructing two separate matrices, we propose a novel and effective way to construct the affinity matrix that encodes distance information from a point to all its important nearest and farthest neighbours points. The affinity matrix is included in manifold learning process for the MTRD data.

Whereas manifold learning has received much attention in MTRD clustering, the effect of interactions among clusters of different object types in MTRD clustering is often neglected. The concept of association relationship was first proposed in [70], therein by formulating the NMF with orthogonal constraint as a preparation for the spectral clustering approach, the association relationship type has been embedded naturally. However, since the optimization has been solved by the spectral approach, the association relationship is not really utilized in the learning process. In BiOR-NM3F [29] through enforcing the orthogonal constraint on factor matrix, the authors conjecture that the trade-off matrix in the problem of tri-factorization becomes association relationship but failed to provide an explanation and analysis. Authors of IGNMTF [42] pointed out the issues such as trivial solutions and bad scales when manifold learning is combined with the factorization process. They proposed a trade-off matrix that supposedly approximates an association relationship matrix between two object types to solve these embedded problems. Yet again the method only mentioned the association relationship type, it offered no formal definition and analysis of its importance to the clustering process.

3.3 The Association Relationship and accurate structure Preserving (ARASP) method

We explain how the proposed ARASP method will be able to learn the association relationships of clusters as well as preserve the close and far distances between data points.

3.3.1 Problem definitions and Notations

Let $\mathcal{D} = \{X_1, X_2, ..., X_m\}$ be a dataset with m object types. Each object type $X_h = \{x_i\}_{(1 \leq i \leq n_h)}$ is a collection of n_h objects. Let $\{R_{hl} \in \mathbb{R}_+\}$ be a set of inter-type relationship matrices where $R_{hl} = \{r_{ij}\}^{n_h \times n_l}$, r_{ij} denotes the inter-type relationship (e.g., the *tf-idf* weight of a term in a document) between ith object and jth object of X_h and X_l, respectively, $R_{lh}^T = R_{hl}$. Let $\{W_h\}_{1 \leq h \leq m}$, a set of intra-type relationship matrices of m object types where $W_h = \{w_{ij}\}^{n_h \times n_h}$ is a weighted adjacency matrix resulted from building a nearest neighbour graph (kNN graph) of object type as in [21, 48, 99].

The task in this paper is to simultaneously cluster m object types into c clusters by considering the intra-type relationships between objects in the same type, the inter-type relationships between objects of different types and the association relationships between clusters. We assume that the different type objects in the same dataset tend to have the same number of underlying groups due to their highly interrelatedness. The proposed method can be applied with minor variations in the case when the number of clusters generated are different among different data types.

The NMF-based objective function combined with manifold learning for MTRD can be written as [99, 70],

$$J_1 = \min \Big(\sum_{1 \leq h < l \leq m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2 + \lambda \sum_{1 \leq h \leq m} Tr(G_h^T L_h G_h) \Big), \text{ s.t., } G_h \geq 0, G_l \geq 0 \tag{3.1}$$

Where $\|.\|_F$ denotes the Frobenius norm and G_h and G_l are low-rank representations of object types X_h and X_l respectively. S_{hl} is a trade-off matrix that provides additional degree of freedom for the factorization process [29]. L_h is the graph Laplacian of object type X_h and is defined as $L_h = D_h - W_h$. D_h is the diagonal matrix computed by $(D_h)_{ii} = \sum_j (W_h)_{ij}$ [99, 21]. λ is playing role as the trade-off parameter to control the balance during optimizing process for the two terms in

objective function J_1.

The first term in Eq. (3.1), called as $P1$, is for learning low dimensional representations and the second term, named as $P2$, is for preserving local geometric structure of data in the newly found low dimensional space. $P2$ can be written as,

$$P_2 = \min \lambda \sum_{1 \leq h \leq m} \big(Tr(G_h^T D_h G_h) - Tr(G_h^T W_h G_h)\big), G_h \geq 0 \tag{3.2}$$

In this paper, we aim to reformulate both components $P1$ and $P2$ such that $P1$ not only learns low-dimensional representations but also learns the association relationships between clusters of different object types and $P2$ not only preserves the local distances between points but also maintains the global geometric structure of the data.

3.3.2 ARASP Objective function formulation

Learning association relationship

In this section, we present the reformulation of $P1$ in objective function $J1$ such that the objective function can learn and control the interactions between clusters of different object types when learning higher order to lower order mapping. An illustration of association relationships between clusters is shown in Figure 3.2. The association relationships between clusters will be learnt by enforcing the Normalize Cut-type constraint [93] on factor matrices. We apply $l1$-normalization constraint on the learnt association relationship with the assumption that objects belonging to a cluster from a data type should have more interaction with a particular cluster than other clusters from the other data type. Next, we formally define the association relationship amongst clusters.

Definition 1: Association Relationship

The association relationship represents the relationship between two clusters

from two different object types. The higher association relationship reflects the more related objects between two clusters.

Formally, let R_{hl} be the inter-type relationships between object type X_h and X_l. Suppose G_h and G_l are the optimal low order representations of X_h and X_l, respectively, by minimizing $P1$. The association relationship matrix $A_{hl} = \{a_{tf}\}^{c\times c}$ that shows relations between clusters of these two object types can be written as,

$$A_{hl} = G_h^T R_{hl} G_l \tag{3.3}$$

Note that $A_{hl} \geq 0$ due to $R_{hl} \geq 0$ and the non-negative constraints on G_h, G_l. This association relationship matrix A_{hl} is denoted as follows,

$$A_{hl} = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1c} \\ a_{21} & a_{22} & ... & a_{2c} \\ ... & ... & ... & ... \\ a_{c1} & a_{c2} & ... & a_{cc} \end{bmatrix} \tag{3.4}$$

where each element a_{tf} indicates the association relationship between (objects of) cluster t of data type X_h and (objects of) cluster f of data type X_l. More specifically, a_{tf} is represented as,

$$a_{tf} = g_t^{h'} R^{hl} g_f^l = \sum_{i=1}^{n_h} \sum_{j=1}^{n_l} g_{it}^h r_{ij}^{hl} g_{jf}^l \tag{3.5}$$

where $g_t^h = [g_{1t}^h, g_{2t}^h, ..., g_{n_h t}^h]' \in G_h$ and $g_f^l = [g_{1f}^l, g_{2f}^l, ..., g_{n_l f}^l]' \in G_l$, the higher the value of a_{tf}, the higher the relatedness between objects of the two clusters t and f of data types X_h and X_l, respectively.

Next we justify how to incorporate learning association relationships into the objective function that can lead to a more meaningful and effective model. As mentioned in Section 3.3.1, S_{hl} in P_1 is the trade-off matrix providing additional

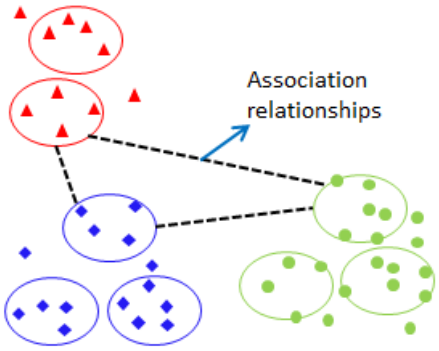


Figure 3.2: Illustration of association relationships between clusters of different data types (the black dotted lines). Red circles show different clusters of data type X_1, blue circles show different clusters of data type X_2, green circles show different clusters of data type X_3.

degree of freedom to the factorization process. We propose to modify S_{hl} to become the association relationship matrix amongst clusters of X_h and X_l. To treat S_{hl} as an association relationship matrix, we first replace S_{hl} in the first term P_1 of Eq. (3.1) by A_{hl} as defined in Eq. (3.3). We have the following equivalent equation,

$$P_1 = \min \sum_{1 \leq h < l \leq m} \|R_{hl} - G_h G_h^T R_{hl} G_l G_l^T\|_F^2 \tag{3.6}$$

The optimization problem in Eq. (3.6) will be optimal when $R_{hl} = G_h G_h^T R_{hl} G_l^T G_l$ where $G_h G_h^T = I$ and $G_l G_l^T = I$. Therefore, we propose that by enforcing orthogonal constraint on factor matrices $G_h, G_l, 1 \leq h < l \leq m$ in P_1, S_{hl} becomes the association relationship matrix A_{hl}. The non-negative and orthogonal constraints on $G_h, G_l, 1 \leq h < l \leq m$ are very strict and can result in a very sparse and poor new representations [29], therefore we relax the orthogonal constraint by using the Normalize Cut-type constraint [93], $G_h^T D_h G_h = I$, $G_l^T D_l G_l = I$ as in [42, 100]. The constraint not only helps relaxing the strict orthogonal constraint but also allows a non-trivial solution [42] with faster convergence [100] due to its benefit of binding the manifold learning regularization to become a non-negative relaxed Normalized

Cut objective function [42]. $P1$ can be rewritten as,

$$\begin{aligned} P_1 &= \min \sum_{1 \le h < l \le m} \|R_{hl} - G_h A_{hl} G_l^T\|_F^2, \\ s.t.\ & G_h \ge 0, G_l \ge 0, G_h^T D_h G_h = I, G_l^T D_l G_l = I, \\ & A_{hl} \ge 0, 1 \le h < l \le m \end{aligned} \tag{3.7}$$

The optimization problem in Eq. (3.7) has incorporated association relationships amongst clusters. We conjecture that by transforming the trade-off matrix S_{hl} to the association relationship matrix A_{hl}, the corresponding factor matrices for object types X_h and X_l will approach to the optimal solutions, i.e., the cluster labels for X_h and X_l will be more meaningful.

We propose applying $l1-$normalization on the rows of association relationships A_{hl} to encourage sparsity [91]. A vector $a_i = [a_{i1}, a_{i2}, ..., a_{ic}]$ from matrix A_{hl} in Eq. (3.4) represents how objects belonging to cluster i of data type X_h are interacting with the objects in different clusters of data type X_l. The highest value of a_{ij} shows that cluster i (X_h) has the highest relatedness with cluster j (X_l). We propose that instead of leaving the values on each vector of A_{hl} uncontrolled, applying $l1-$normalization will help to control the association relationships of a cluster (from X_h) to different clusters of X_l by rewarding the association relationships between related clusters and by penalizing the association relationships between irrelevant clusters. In this way the learning process will bring similar objects closer to help resulting in a meaningful clustering.

Learning accurate structure of data

In this section, we present the reformulation of $P2$ in objective function $J1$ such that the objective function can learn and preserve the accurate geometric structure of the original space in the new space. To learn the original geometric structure of data, we propose to construct the affinity matrix W_h in a novel fashion such that it

can capture closeness and at the same time ensure the looseness between data points of the same data type, X_h. W_h includes the combination of a weighted Adjacency matrix W_h^n and Repulsive matrix W_h^f.

Suppose $T = \{t_{ij}\}^{n_h \times n_h}$ represents the similarity between all pairs of data points x_i and x_j of data type X_h. Different similarity measures can be used such as Binary, Heat Kernel, Cosine or Dot-Product [21, 44].

The weighted Adjacency matrix W_h^n of data type X_h is constructed by building a k nearest neighbour (kNN) graph to model the local geometric structure of data. $W_h^n = \{w_h^n(i,j)\}^{n_h \times n_h}$ is defined as follows,

$$w_h^n(i,j) = \begin{cases} t_{ij} \text{ if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{3.8}$$

where $\mathcal{N}_k(x_i)$ denotes the k nearest neighbour points of x_i.

W_h^n is built with the goal of preserving the local affinity structures of data when projecting to lower order space.

The weighted Repulsive matrix is constructed by building a p farthest neighbour (pFN) graph to encode the discriminative information of data space. $W_h^f = \{w_h^f(i,j)\}^{n_h \times n_h}$ is defined as follows,

$$w_h^f(i,j) = \begin{cases} t_{ij} \text{ if } x_i \in \mathcal{F}_p(x_j) \text{ or } x_j \in \mathcal{F}_p(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{3.9}$$

where $\mathcal{F}_p(x_i)$ denotes the p farthest neighbour points of x_i.

W_h^f is built with the goal of keeping non-similar points far apart in the new space since there is no chance these points will share the same clusters.

For the purpose of keeping similar points close and dissimilar points distant, the affinity matrix $W_h = \{w_h(i,j)\}^{n_h \times n_h}$ for data type X_h will be constructed by linearly

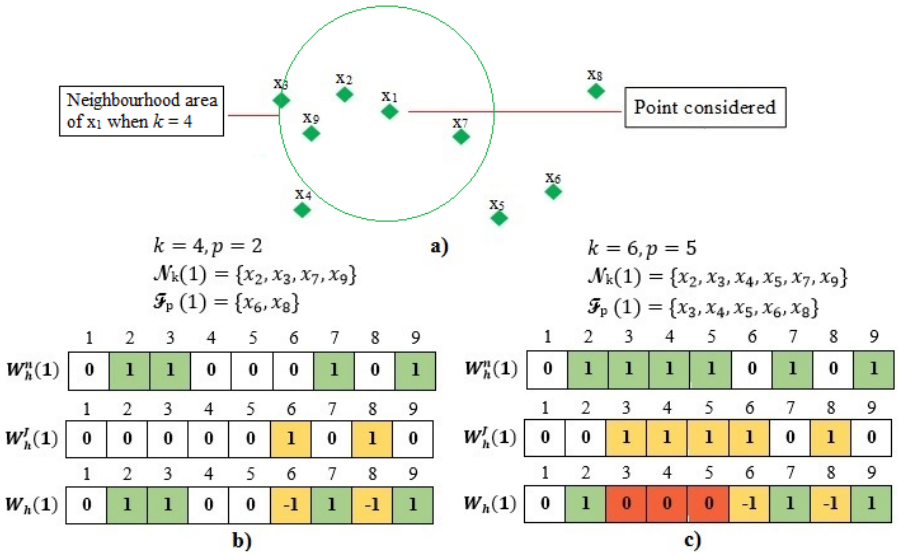


Figure 3.3: Construction of the new affinity weight matrix based on Eq. (3.10). For the simplicity, we use Binary weight and consider the point x_1 only. Affinity between point x_1 to all its k nearest neighbours and p farthest points are given in $W_h^n(1,j)$ and $W_h^f(1,j), j = 1..n_h$, respectively. Fig 3.3.a illustrates data points of data type X_h lying on R^2. Fig 3.3.b illustrates W_h^n, W_h^f and W_h when $k = 4, p = 2$. Fig 3.3.c illustrates W_h^n, W_h^f and W_h when $k = 6, p = 4$.

combining two matrices W_h^n and W_h^f as follows,

$$W_h = \alpha W_h^n - \beta W_h^f, \alpha + \beta = 1 \tag{3.10}$$

A possibility of overlapped points, i.e., a point appears in both the nearest and farthest sets, will occur when $w_h^n(i,j) \neq 0$, $w_h^f(i,j) \neq 0$, i.e., x_j is residing along the border area where it may belong to the nearest neighbourhood or farthest neighbourhood of x_i. In this case, we treat the point x_j as neutral and assign a value of 0 for $w_h(i,j)$. More specifically, $w_h(i,j) = 0$ if $w_h^n(i,j) \neq 0$ and $w_h^f(i,j) \neq 0, \forall x_i, x_j \in X_h$. Figure 3.3.c illustrates this case where x_3, x_4, x_5 are treated as neutral points.

The proposed approach of constructing W_h allows including both close and far distances of original data. It remains proving that using this new W_h will maintain the distances in new mapped space, i.e., close points and far points in original space will be kept close and far in new space, respectively. We have the following lemma,

Lemma 1: Use of the newly constructed W_h in learning will ensure that the

factorization process maintains the close and far distances between points.

Proof: From the graph regularization term $P2$ in Eq. (3.2) that learns embedded manifolds on all data types, we have the following rewritten formulation of graph regularization to learn embedded manifold on data type X_h [10, 21].

$$Tr(G_h^T D_h G_h) - Tr(G_h^T W_h G_h) = \frac{1}{2}\sum_{i,j=1}^{n_h} \|g_i - g_j\|^2 w_h(i,j) \tag{3.11}$$

where $w_h(i,j)$ is the affinity between two points x_i and x_j of object type X_h. $w_h(i,j)$ constructed as in Eq. (3.10), will belong to one of the three cases.

- If $w_h(i,j) > 0$, i.e., x_i and x_j are close in original space, minimizing Eq. (3.11) will lead to minimizing $\sum_{i,j=1}^{n_h} \|g_i - g_j\|^2$, or minimizing the distance between two new representations g_i, g_j of x_i, x_j, i.e., the new representations g_i and g_j will remain close in new space.

- If $w_h(i,j) < 0$, i.e., x_i and x_j are far in original space, minimizing Eq. (3.11) will lead to maximizing $\sum_{i,j=1}^{n_h} \|g_i - g_j\|^2$, or maximizing the distance between two new representations g_i, g_j of x_i, x_j, i.e., the new representations g_i and g_j will remain distant in new space.

- If $w_h(i,j) = 0$, i.e., x_i and x_j have no interdependency (they are neither in near neighbourhood nor in far neighbourhood areas) in original space, the minimization process in Eq. (3.11) will ignore these points.

This lemma leads us to infer that using the new construction of affinity matrix will simultaneously maintain the distances from an object to its close and remote objects and thus help finding the low dimensional representation of data, projecting the accurate and complete structure of data. One may argue that the far away distances in the ambient space may not reflect the true non-linear structure. However, if learning process ignores far points, i.e., far points have not been kept far apart, the far points will be treated as relevant points (especially when choosing a wrong nearest neighbourhood size k). Thus may lead to a new low-order data that does

not respect the original structure.

Adopting the novel fashion of constructing W_h as in Eq. (3.10), we construct $\{W_h\}_{(1\leq h\leq m)}$ for all m object types and substitute for W_h in Eq. (3.2) to have a new underlying meaning for term $P2$ and thus learning accurate embedded manifold on each data type.

Combining Eq. (3.2) and (3.7) with the proposed $\{W_h\}_{1\leq h\leq m}$ as in Eq. (3.10), the proposed meaningful MTRD clustering problem can be defined as the following minimization problem,

$$\begin{aligned} J_1 = &\min \Big(\sum_{1\leq h<l\leq m} \|R_{hl} - G_h A_{hl} G_l^T\|_F^2 + \sum_{1\leq h\leq m} \big(Tr(G_h^T D_h G_h) - Tr(G_h^T W_h G_h)\big)\Big), \\ &\text{s.t., } G_h^T D_h G_h = I, G_l^T D_l G_l = I, G_h \geq 0, G_l \geq 0, A_{hl} \geq 0, 1 \leq h < l \leq m \end{aligned} \tag{3.12}$$

The $\{D_h\}_{(1\leq h\leq m)}$ in Eq. (3.12) is constructed based on new constructed $\{W_h\}_{(1\leq h\leq m)}$.

In this objective function, we remove the Laplacian parameter λ used in Eq. (3.2) as we already have parameters α and β when constructing W_h that can be adjusted in different cases. Notice that due to the added Normalize Cut-type constraint, $Tr(G_h^T D_h G_h)$, $Tr(G_l^T D_l G_l)$, $\forall h, l, 1 \leq h < l \leq m$ are constant, Eq. (3.12) is equivalent to,

$$\begin{aligned} J_1 = \min \Big(&\sum_{1\leq h<l\leq m} \|R_{hl} - G_h A_{hl} G_l^T\|_F^2 \\ &- \sum_{1\leq h\leq m} Tr(G_h^T W_h G_h)\Big), \\ G_h \geq 0, G_l \geq 0, A_{hl} \geq 0, &G_h^T D_h G_h = I, G_l^T D_l G_l = I, \\ G_h 1_c = 1_{n_h}, G_l 1_c = 1_{n_l}, &A_{hl} 1_c = 1_c, 1 \leq h < l \leq m \end{aligned} \tag{3.13}$$

where $1_c, 1_{n_h}, 1_{n_l}$ are column vectors with sizes c, n_h, n_l, respectively, with all elements are 1s. The $l1-$normalization is applied on every row of G_h, G_l to ensure the optimization problem in Eq. (3.13) is well defined and no longer suffers from trivial

solution [43, 48].

The proposed objective function with the novel construction of W_h, association relationship A_{hl} and non-negative and Normalize Cut-type constraints will learn accurate low-representations by including association relationships and maintaining the local and global geometric structures of data. The derived optimization problems of clustering can be generalized to traditional "flat" data clustering and co-clustering straightforwardly by allowing the number of object types m to be 1 or 2, respectively in Eq. (3.13).

3.3.3 Algorithmic Solution to the ARASP

In this section, we provide the detail of the solution for the proposed objective function in Eq. (3.13) with respect to A_{hl}, G_h, G_l. We will separately update each variable while fixing others as constants until convergence [43, 66, 99] and introduce the iterative algorithm 1 to simultaneously group objects of several different types into clusters.

1. Solving A_{hl}: When fixing h, l and thus fixing G_h, G_l, Eq. (3.13) is reduced to minimizing

$$J_{A_{hl}} = \|R_{hl} - G_h A_{hl} G_l^T\|_F^2, A_{hl} \geq 0 \tag{3.14}$$

$$\partial J_{A_{hl}}/\partial A_{hl} = -2R_{hl}G_h^T G_l + 2A_{hl}G_l^T G_l G_h^T G_h \tag{3.15}$$

$$\Leftrightarrow 2R_{hl}G_h^T G_l = 2A_{hl}G_l^T G_l G_h^T G_h \tag{3.16}$$

Following the similar condition in [29] for non-negativity of A_{hl}, we have the update rule for A_{hl}

$$A_{hl} = (A_{hl})_{ij}\left[\frac{(G_h^T R_{hl} G_l)_{ij}}{(G_h^T G_h A_{hl} G_l^T G_l)_{ij}}\right]^{1/2} \tag{3.17}$$

2. Solving G_h, G_l:

Solving $G_h, G_l, 1 \leq h < l \leq m$ is obviously equivalent to optimize all G_h where $1 \leq h \leq m$. When fixing h, fixing $A_{hl}, G_l, h < l \leq m$, fixing $A_{lh}, G_l, 1 \leq l < h$, optimizing Eq. (3.13) with respect to G_h is equivalent to optimizing the following,

$$\begin{aligned} J_{G_h} = & \sum_{h<l\leq m} \|R_{hl} - G_h A_{hl} G_l^T\|_F^2 + \sum_{1\leq l<h} \|R_{lh} - G_l A_{lh} G_h^T\|_F^2 - Tr(G_h^T W_h G_h), \\ & \text{s.t., } G_h \geq 0, G_h^T D_h G_h = I \end{aligned} \tag{3.18}$$

Since $G_h^T D_h G_h = I$, we introduce the Lagrangian multiplier Λ, thus the Lagrangian function becomes,

$$\begin{aligned} L_{G_h} = & \sum_{h<l\leq m} \|R_{hl} - G_h A_{hl} G_l^T\|_F^2 + \sum_{1\leq l<h} \|R_{lh} - G_l A_{lh} G_h^T\|_F^2 \\ & -Tr(G_h^T W_h G_h) + Tr(\Lambda(G_h^T D_h G_h - I)) \end{aligned} \tag{3.19}$$

By taking the derivative of L_{G_h} on G_h we have:

$$\begin{aligned} \partial L_{G_h}/\partial G_h = & \sum_{h<l\leq m} (-2R_{hl} G_l A_{hl}^T + 2G_h A_{hl} G_l^T G_l A_{hl}^T) \\ & + \sum_{1\leq l<h} (-2R_{lh}^T G_l A_{lh} + 2G_h A_{lh}^T G_l^T G_l A_{lh}) - 2W_h G_h + 2D_h G_h \Lambda \end{aligned} \tag{3.20}$$

Let

$$P = \sum_{h<l\leq m} (R_{hl} G_l A_{hl}^T) + \sum_{1\leq l<h} (R_{lh}^T G_l A_{lh}) \tag{3.21}$$

$$Q = \sum_{h<l\leq m} (A_{hl} G_l^T G_l A_{hl}^T) + \sum_{1\leq l<h} (A_{lh}^T G_l^T G_l A_{lh}) \tag{3.22}$$

Since the Karush-Kuhn-Tucker (KKT) condition [18] for the non-negative con-

Algorithm 1: Association Relationship and Accurate Structure Preserving

Input : Inter relationship matrices $\{R_{hl}\}_{1\leq h<l\leq m}^{n_h\times n_l}$, clusters number c, parameters α, β, k and p with non-negative values as discussed in the previous section.

Output: Cluster indicator matrices: $G_h, 1 \leq h \leq m$

Initialize non-negative matrices $\{G_h\}_{1\leq h\leq m}$ by K-means, using inter relationship matrices $\{R_{hl}\}_{1\leq h<l\leq m}^{n_h\times n_l}$ and clusters number c as input.

for *each* $h, 1 \leq h \leq m$ **do**
  compute the similar weighted Adjacency matrix W_h^n, Repulsive weighted matrix W_h^f and the combined matrix W_h as in Eqs. (3.8)-(3.10).
end
repeat
  for *each* $(h, l), 1 \leq h < l \leq m$ **do**
    Update A_{hl} as in Eq. (3.17)
    Normalize A_{hl}
  end
  for *each* $h, 1 \leq h \leq m$ **do**
    Update G_h as in Eq. (3.24)
    Normalize G_h
  end
until *converges*;
Transform each $\{G_h\}_{1\leq h\leq m}$ into cluster indicator matrices by K-means.

straint on G_h gives $(\frac{\partial L_{G_h}}{\partial G_h})_{ij}(G_h)_{ij} = 0$, we have the following,

$$\Lambda = 2G_h^T P - 2G_h^T G_h Q + 2G_h^T W_h G_h \tag{3.23}$$

Since Λ, W_h and D_h may take mixed signs, we introduce $\Lambda = \Lambda^+ - \Lambda^-$, where $\Lambda_{ij}^+ = (|\Lambda_{ij}^+| + \Lambda_{ij}^+)/2$, $\Lambda_{ij}^- = (|\Lambda_{ij}^+| - \Lambda_{ij}^+)/2$ and similarly to $W_h^{+/-}$, $D_h^{+/-}$ [30]. The update rule for G_h becomes,

$$(G_h)_{ij} = (G_h)_{ij}\left[\frac{(P + W_h^+ G_h + D_h^- G_h \Lambda^-)_{ij}}{(G_h Q + W_h^- G_h + D_h^+ G_h \Lambda^+)_{ij}}\right]^{1/2} \tag{3.24}$$

Algorithm 1 summarizes the proposed ARASP approach. By iteratively using update rule for A_{hl} as in Eq. (3.17) that applies non-negative and $l1-$normalization and update rule for G_h as in Eq. (3.24) that applies pseudo-orthogonal constraint and incorporates local and global structures of data, ARASP learns not only an

Table 3.1: Characteristics of the datasets

Properties	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
# Classes	25	10	5	7	17	4
# Object types	3	3	3	3	3	2
# Samples	1,413	1,500	187	2,708	617	9,625
# Terms, # Actors	2,921	3,000	1,702	1,433	1,398	29,992
# Concepts, # Links, # Keywords	2,437	3,000	578	5,429	1,878	-

accurate shape of data but also the association relationships between clusters of different object types, to provide a meaningful and accurate clustering results.

3.3.4 Convergence Analysis of ARASP

We prove the convergence of ARASP algorithm through update rules of A_{hl}, G_h as shown in Eqs. (3.17) and (3.24) respectively by using the auxiliary function approach [63].

Definition 3: $Z(h, h')$ is an auxiliary function for $F(h)$ if the conditions $Z(h, h') \geq F(h)$ and $Z(h, h) = F(h)$, are satisfied. [63]

Lemma 2: If Z is an auxiliary function for F, then F is non-increasing under the update [63]

$$h^{(t+1)} = arg \min_{h} Z(h, h^{(t)})$$

Proof:

$F(h^{(t+1)} \leq Z(h^{(t+1)}, h^{(t)}) \leq Z(h^{(t)}, h^{(t)}) = F(h^{(t)})$

Lemma 3: For any non-negative matrices $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{k \times k}, S \in \mathbb{R}^{n \times k}, S' \in \mathbb{R}^{n \times k}$, when A and B are symmetric, the following inequality holds [30]

$$\sum_{i=1}^{n} \sum_{p=1}^{k} \frac{(AS'B)_{ip} S_{ip}^2}{S'_{ip}} \geq Tr(S^T ASB)$$

Next, we will present two theorems that ensure the convergence of ARASP relating to the use of auxiliary function.

Theorem 1: Let $L(G_h) = Tr(-G_h^T P + G_h^T Q G_h - G_h^T W_h G_h + G_h^T D_h G_h \Lambda)$

$$L(G_h) = Tr(-G_h^T P + G_h^T Q G_h - G_h^T W_h^+ G_h + G_h^T W_h^- G_h + G_h^T D_h^+ G_h \Lambda^+ - G_h^T D_h^- G_h \Lambda^-)$$

Then the following function

$$Z(G_h, G_h') = -\sum_{ij} P_{ij}(G_h')_{ij}(1 + log\frac{(G_h)_{ij}}{(G_h')_{ij}}) + \sum_{ij} \frac{(G_h' Q)_{ij}(G_h^2)_{ij}}{(G_h')_{ij}}$$
$$-\sum_{ijk}(W_h^+)_{jk}(G_h')_{ji}(G_h')_{ki}(1 + log\frac{(G_h)_{ji}(G_h)_{ki}}{(G_h')_{ji}(G_h')_{ki}})$$
$$+\sum_{ij} \frac{(G_h' W_h^-)_{ij}(G_h^2)_{ij}}{(G_h')_{ij}} + \sum_{ij} \frac{(D_h G_h' \Lambda^+)_{ij}(G_h^2)_{ij}}{(G_h')_{ij}}$$
$$-\sum_{ijkl}(\Lambda^-)_{kj}(D_h)_{jl}(G_h')_{ji}(G_h')_{lk}(1 + log\frac{(G_h)_{ji}(G_h)_{lk}}{(G_h')_{ji}(G_h')_{lk}})$$

is an auxiliary function for $L(G_h)$. Furthermore, it is a convex function in G_h and its global minimum is the update rule of G_h given in Eq. (3.24).

Proof: Using Lemma 3 and the inequality $z \geq (1 + logz), \forall z > 0$, Theorem 1 can be proved straightforwardly.

Theorem 2: Updating $G_h, 1 \leq h \leq m$ using the update rule as in Eq. (3.24) will monotonically decrease the objective function as in Eq. (3.13).

Proof: According to Lemma 2 and Theorem 1, we have

$$L(G_h^0) = Z(G_h^0, G_h^0) \geq Z(G_h^1, G_h^0) \geq L(G_h^1) \geq ...$$

Therefore $L(G_h)$ is monotonically decreasing.

Theorems 1 and 2 guarantee the convergence of the ARASP Algorithm regarding to G_h. The sub problem in Eq. (3.14) is a convex problem regarding to A_{hl} with the global minima is in Eq. (3.17). The correctness of the algorithm is also guaranteed as the update rules in Eq. (3.24) satisfies the Karush-Kuhn-Tucker optimal condition [18]. Thus we conclude that, the objective function of ARASP as defined in Eq. (3.13) monotonically decreases and guarantees convergence in finite time to an optimal solution.

3.3.5 Complexity Analysis

Time complexity of ARASP is composed of two parts: (1) learning the accurate manifold and (2) applying multiplicative update rules.

Table 3.2: NMI for each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
NMF	66.17	50.40	14.68	11.55	29.22	39.94	35.33
DRCC-Extended	72.71	59.68	35.78	31.66	**33.93**	37.88	45.27
STNMF	67.28	57.89	31.35	20.20	30.93	39.93	41.26
RHCHME	70.36	62.04	33.15	33.53	31.34	50.46	46.81
MultiOR-NM3F	61.82	63.41	23.92	31.46	27.71	45.27	42.30
ARP	71.64	67.63	35.28	33.61	32.91	49.19	48.38
ARASP	**75.01**	**69.33**	**39.23**	**35.24**	33.18	**51.28**	**50.55**

Table 3.3: Accuracy for each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
NMF	57.61	58.87	59.36	30.21	28.53	61.43	49.34
DRCC-Extended	57.32	59.13	62.57	50.33	32.58	66.96	54.82
STNMF	51.38	63.87	44.92	41.29	26.58	59.92	47.99
RHCHME	65.18	64.13	50.27	55.28	31.44	72.98	56.55
MultiOR-NM3F	49.47	67.33	47.06	55.76	28.69	70.87	53.20
ARP	57.61	71.80	53.48	57.90	30.47	70.97	57.04
ARASP	**71.05**	**73.33**	**59.36**	**58.31**	**33.39**	**74.22**	**61.61**

To learn the accurate manifold in ARASP, the affinity matrix W_h of object type X_h is constructed based on both kNN and pFN graphs thus contains $k+p$ non-zero elements on each row, encoding the closeness information to k nearest neighbours and p farthest neighbours for each point. This step requires $O(n_h^2(k+p))$ and thus requiring $O(mn_h^2(k+p))$ to construct the affinity matrices for all m object types.

The cost of applying multiplicative update rules includes $O(n_h n_l c\frac{m(m-1)}{2})$ to learn the association relationship matrices $\{A_{hl}\}_{1\leq h<l\leq m}$ and $O(n_h^2 cm(m-1))$ for updating $\{G_h\}_{1\leq h\leq m}$ therefore generates a time complexity of $O(m(m-1)n_h^2 c)$ for the step of multiplicative updating. The overall time complexity of ARASP is $O(mn_h^2(k+p)) + O(m(m-1)n_h^2 c)$, where m is the number of data types, n_h, n_l are the number of objects in object types X_h, X_l and c is the number of clusters, respectively.

The other benchmark method, e.g., DRCC-Extended consumes an overall time cost of $O(mn_h^2 k) + O(m(m-1)n_h^2 c)$ to learn the manifold and to apply updating wherein the affinity matrices are built based on kNN graph only and the updating

factor matrices ignores learning association relationships.

Size of every object type is about the same, and m and c are extremely smaller than the object type size, the computational complexity of the proposed algorithm remains as quadratic. This is similar to the existing MTRD algorithms. This shows that the embedded steps to learn the accurate manifold and the association relationship will not incur the extra cost, whereas improve the accuracy.

3.4 Experiments

3.4.1 Datasets

We evaluate the proposed method on several real world datasets, available with ground truth, that have been commonly used in existing multi-type relational data clustering methods. They are R-MinMax (D1), R-Top (D2), Texas (D3), Cora (D4), Movie (D5), and RCV1-4Class (D6) as detailed in Table 3.1.

R-MinMax (D1) and R-Top (D2) are two subsets of Reuters-21578[1], a very well-known text dataset. To create MTRD, we used external knowledge i.e., Wikipedia, and follow the process as in [48, 57] to generate concepts as the third data type (along with documents and terms data types). The processed datasets present three inter relationships in the form of document-term, document-concept and term-concept co-occurrence matrices and three affinity matrices derived using Eq. (3.10). The associated clustering task on datasets D1 and D2 is to simultaneously cluster all object types in the datasets to enhance the performance of clustering documents.

Texas[2] (D3) dataset describes web pages of Texas University, representing three data types: web pages, words and links between web pages. Clustering task is to achieve different groups of web pages. Cora[3] (D4) is a scientific publication dataset

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] http://www.cs.umd.edu/projects/linqs/projects/lbc/
[3] http://www.cs.umd.edu/%7Esen/lbc-proj/data/cora.tgz

that includes object types publications along with links and words. The Movie dataset (D5) is extracted from the IMDB Website[4] that groups movies into genres by simultaneously clustering three object types: movies, actors and keywords. We also check the effectiveness of ARASP by testing it on a bi-type dataset, RCV1-4Class (D6), a subset of RCV1 corpus [65] that only contains documents and terms.

3.4.2 Evaluation Criteria and Benchmark methods

We utilize the two widely applied measurement criteria, clustering accuracy (AC), the percentage of correctly obtained labels and Normalized Mutual Information (NMI) [119]. We also examine the computational time of our method with benchmark methods. Average results are produced after 5-fold runs of each experiment. ARASP is compared with the following traditional, co-clustering and MTRD state-of-the-art clustering methods.

- DRCC [43] is a co-clustering method of manifold learning on both data samples and features. Since DRCC is a bi-type clustering method, we extended it to have **DRCC-Extended**, that can cluster many related object types simultaneously.

- STNMF [99] simultaneously clusters multi-type data using inter-type relationships and includes intra-type relationships through graph regularization by formulating a new symmetric form of the factor matrix.

- RHCHME [48] also uses symmetric form for the factor matrix and attempts to learn the complete and accurate intra-type relationships by considering data lying on manifolds as well as on subspaces.

- BiOR-NM3F [29] is the first work that applies the orthogonal constraint on factor matrices of both samples and features leading to a well-defined NMF co-

[4] http://www.imdb.org

clustering based method. We extend this method so it can perform clustering on multi-type data, named as **MultiOR-NM3F**.

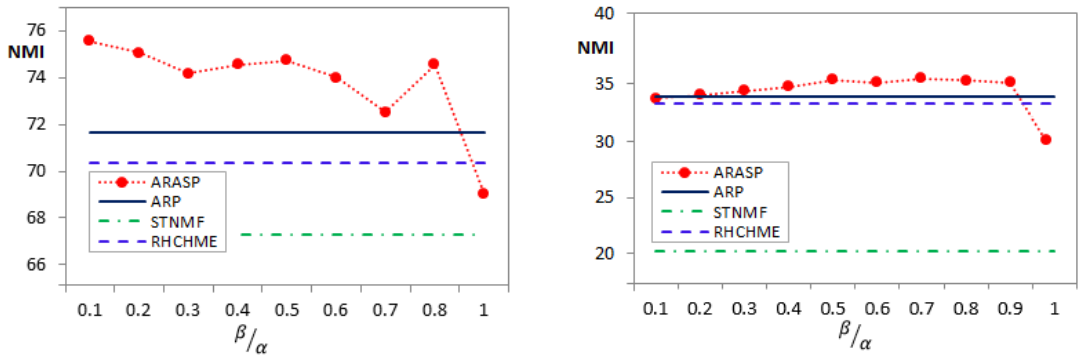


Figure 3.4: NMI changes with the alterations of $\frac{\beta}{\alpha}$ on datasets D1 and D4

- To check the important role of learning an accurate structure when representing data, we implement a version of ARASP which is referred to as **ARP** (Association Relationship Preserving) applied on MTRD where the learning process maintain the association relationship and close distances between data points and ignore ensuring far distances between data points.

- In addition, we selected conventional clustering method NMF [64] to evaluate the effect of MTRD learning as it exploits more information.

  Apart from being applied naturally on the dataset D6 to cluster documents, NMF uses co-occurrences between documents and terms on D1-2; relationships between web pages and words on D3; between publications and words on D4; and between movies and actors on dataset D5. Other MTRD methods are applied on MTRD datasets naturally and consider the dataset D6 as a two-type dataset.

3.4.3 Clustering results

As shown in Tables 3.2-3.4, ARASP clearly outperforms the state-of-art benchmarking methods on all datasets due to its capability to utilize available information as

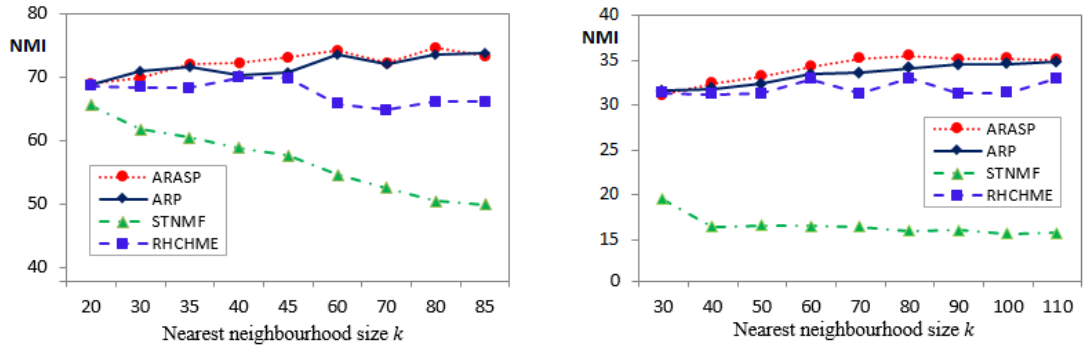


Figure 3.5: NMI changes with the alterations of nearest neighbourhood size k on datasets D1 and D4

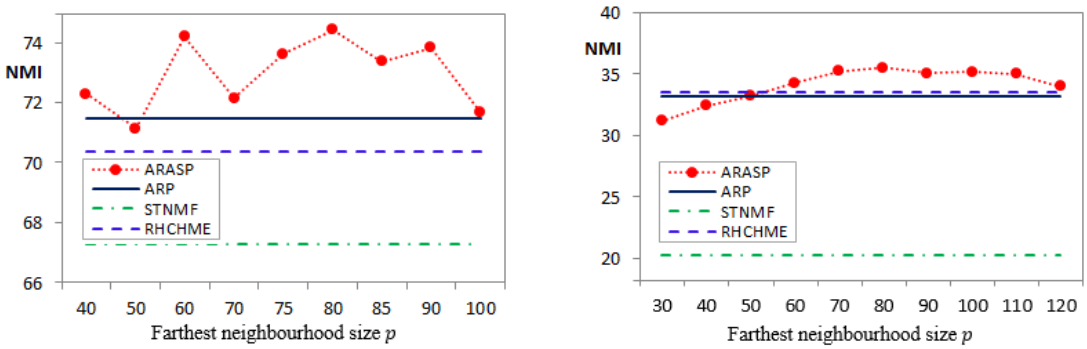


Figure 3.6: NMI changes with the alterations of farthest neighbourhood size p on datasets D1 and D4

much as possible.

First, similar to other NMF-based methods, ARASP can learn low order representations of data thus can keep the most important features to be used in the clustering process. Second, by utilising a novel constructed Affinity matrix that encodes not only close distances but also far distances between objects on each object type, ARASP can preserve both local and global geometric structures of data. Third, consideration of association relationships (AR) between clusters enhances the effective learning process. These strong points lead to a stable and more effective clustering method compared to existing methods over a large range of values of parameters.

We also observe that the clustering results are reverted on DRCC-Extended and MultiOR-NM3F, i.e., DRCC-Extended performs better on datasets D1, D3, D5 but worse on datasets D2, D4, D6 than MultiOR-NM3F. The reason lies in

the characteristics of the two methods. While MultiOR-NM3F applies orthogonal constraint (i.e., partial AR), it fails to preserve geometric structure of data, DRCC-Extended does not take AR into consideration but is aided by data local structure learning.

ARP performs mostly superior amongst the benchmarked methods as it enjoys the association relationship and local geometric structure learning. But the proposed ARASP method outperforms it on all datasets as ARP fails to keep far points far when projecting to lower space leading to learning an inaccurate manifold. This shows the effect of using farthest neighbours in learning an accurate manifold and truly justifies the importance of keeping far points distances for global geometric structure.

RHCHME always gives better results compared to STNMF which shares the same framework with RHCHME since the former exploits not only data lying on manifold but data belonging to subspace as well. These empirical results prove the fact that the more information exploited, the higher results will be achieved.

Amongst the six benchmark methods, NMF gives the poor results compared to others since it is the traditional clustering method which cannot simultaneously clustering many object types and fails to preserve both local and global structure of data as well as ignores association relationships between clusters.

With regards to time complexity, ARASP consumes less time than all other MTRD clustering methods (except MultiOR-NM3F). This is evidence of the fast convergence due to imposing the Normalize Cut-type constraint on factor matrices during the learning process. MultiOR-NM3F consumes the least time as it ignores the sampling on manifolds of data and saves time by not computing affinity matrix whereas all methods need to calculate it due to relying on manifold learning. MultiOR-NM3F even converges faster than traditional NMF since it has partial association relationships learning to guide the learning process.

Table 3.4: Running time (in 10^3 seconds) of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
NMF	0.16	0.15	0.015	0.126	0.054	7.54
DRCC-Extended	0.20	0.19	0.014	0.164	0.054	7.28
STNMF	0.22	0.24	0.016	0.136	0.057	8.22
RHCHME	0.22	0.25	0.016	0.139	0.058	8.33
MultiOR-NM3F	0.08	0.06	0.007	0.046	0.037	5.77
ARP	0.14	0.10	0.012	0.073	0.041	5.75
ARASP	0.14	0.10	0.012	0.074	0.040	5.81

3.4.4 Parameters setting

ARASP uses four parameters, i.e., α for the regularization of local structure, β for the regularization of global structure, k is the nearest neighbourhood size and p is the repulsive neighbourhood size. We investigate how the performance of ARASP changes with the alterations of these parameters on datasets D1 and D4.

Since all methods, except MultiOR-NM3F, rely on manifold learning to preserve the local structure of data use the Laplacian regularization parameters α and the neighbourhood size k, we set the same range of values for these parameters, i.e., α is selected from the range $\{1, 10, 20, 50, 100, 200\}$ and k is selected from the range $\{20, 25, 30, 40, 50, 60, 70, 80, 100, 110, 120\}$. To enable a fair comparison between methods, we selected the highest performing set of parameters for each benchmark method on each dataset. Specifically, RHCHME and STNMF reach their peaks when $\alpha = 10$ on D1 and when $\alpha = 20$ on D4. $\alpha = 20$ and $\alpha = 50$ are selected for ARP on datasets D1 and D4, respectively. The k values are depicted in Figure 3.5.

To set parameters α and β in Eq. (3.10), we fix the value for α and change the value for β to allow the rate $\frac{\beta}{\alpha}$ ranges from 0.1 to 1. $\frac{\beta}{\alpha} = 1$ means both components are treated equally. On dataset D1, the local graph regularization α is fixed as 20 and on dataset D4, $\alpha = 100$. ARASP achieves better performance when $\frac{\beta}{\alpha} = [0.1..0.9]$ on the two datasets. This allows us to that (1) the actual value does not conclude matter as long as it is not extreme and (2) preserving distances of far points by

using a pFN graph plays a significant role into the learning process (Figure 3.4).

Figures 3.5 and 3.6 validate the independence of the ARASP on choosing a specific nearest neighbourhood and repulsive neighbourhood sizes k and p respectively. As can be seen from Figure 3.5, benchmark methods utilising kNN graph to capture local structure of data heavily depend on the number of nearest neighbours k. For example, on dataset D1, RHCHME reaches its peak when $k = 40$ and produces fairly lower results with varied k values. Whereas ARASP performs well with a wide range of values for neighbourhood size k, e.g., from 20 to 85 on D1 and from 30 to 110 on D4. Similarly, ARASP does not depend on choosing a specific p, there is no significant change in performance with varied p values (Figure 3.6). None of the existing methods have this component, however their performance is included to show that majority of p values yields higher performance than these methods.

3.5 Conclusion

The paper presents a comprehensive NMF framework for the multi-type relational data and an effective Association relationship and accurate structure preserving (ARASP) clustering method. ARASP aims at finding meaningful clusters by preserving intrinsic accurate structure of data while projecting from higher to lower dimensions, as well as respecting the association relationships between clusters of different data types during the learning process. We specifically present a novel and efficient construction of Affinity matrix embedding both similar and dissimilar information between data points. This newly constructed affinity matrix can also be deployed in the problem of dimensionality reduction. Experimental results on several benchmark datasets show the capacity of ARASP for providing a meaningful clustering solution evidenced by higher accuracy and NMI measures, as well as yielding a lower execution time. Empirical analyses ascertain that ARASP is able to exploit most information available in data and during the clustering process without

compromising the computational cost. In future work, we will investigate how some properties of dataset such as the density, the number of dimensions affect the choice of neighbourhood size. Different normalizations or constraints can be applied on association relationship matrix to estimate the influence of association relationship on learning latent features for data.

Paper 3. T.M.G, Tennakoon, Khanh Luong, Wathsala Mohotti, Sharma Chakravarthy, and Richi Nayak, *Multi-type Relational Data Clustering for Community Detection by Exploiting Content and Structure Information in Social Networks.* The 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019) (In Press).

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors:

T.M.G. Tennakoon (PhD Candidate): Participated in forming the idea of using Multi-type Relational Data approach for the problem of Community Discovery. Collected, preprocess and prepared the datasets. Conducted experiments and analysed experiment results related to structure-based views and community discovery methods Louvain, CNM and InfoMap. Wrote part of the Introduction section, Related Work, part of Proposed Solution and Conclusion and Future work.

Signature: ***Date:*** 24-06-2019

Khanh Luong (PhD Candidate): Participated in forming the idea of using Multi-type Relational Data approach for the problem of Community Discovery. Proposed MTCD (Multi-Type Community Discovery) method to apply in Community Discovery setting; conducted experiments and analysed experiment results for the proposed method MTCD and the benchmarking method MTRD-NMF. Wrote Proposed Method and Solution Section, wrote part of Introduction section and part of Emperical Analysis section.

Signature: ***Date:*** 24-06-2019

Wathsala Mohotti (PhD Candidate): Created the Content-centric Relation. Conducted experiments using content information(tweet text) and non-content information(Urls and Tags) after special short text pre-processing to address the unstructured text in social media and analyzed the results. Benchmarked the content-centric relation against relevant methods(k-means, LDA and NMF). Wrote the Empirical Analysis and part of the Proposed solution related to Content-centric Relation.

Signature: ***Date:*** 24-06-2019

Prof. Sharma Chakravarthy: Provided critical comments on the design and formulation of the concepts, method and experiments.

Signature: ***Date:***

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.

Name: Dr Richi Nayak, Associate Professor

Signature: ***Date:*** 24-06-2019

ABSTRACT. Social Networks popularity has facilitated the providers with an opportunity to target specific user groups for various applications such as viral marketing and customized programs. However, the volume and variety of data present in a network challenge the identification of user communities effectively. The sparseness and heterogeneity in a network make it difficult to group the users with similar interests whereas the high dimensionality and sparseness in text pose difficulty in finding content focused groups. We present this problem of discovering user communities with common interests as the multi-type relational data (MTRD) learning with the content and structural information, and propose a novel solution based on non-negative matrix factorization with added regularization. We empirically evaluate the effectiveness of the proposed method on real-world Twitter datasets benchmarking with the state-of-the-art community discovery and clustering methods.

KEY TERMS. Twitter; Community Discovery; Multi-Type Relational Data Clustering; Non-negative Matrix Factorization;

3.6 Introduction

Social networks such as Twitter have become a popular source of sharing information and opinions. A large volume of data is continuously generated on these networks that create many opportunities for political parties, businesses, and government organizations to target certain audience groups for their campaigns, marketing strategies, and customized programs and events. Community discovery is a well-studied research problem that can facilitate these applications. A large proportion of research has been focused on using graph models, which represent *how* users are connected [17, 40, 86]. However, these structure-based community discovery methods have shown to be ineffective in identifying users with similar interests due to the network sparseness and heterogeneity. On the contrary, clustering users

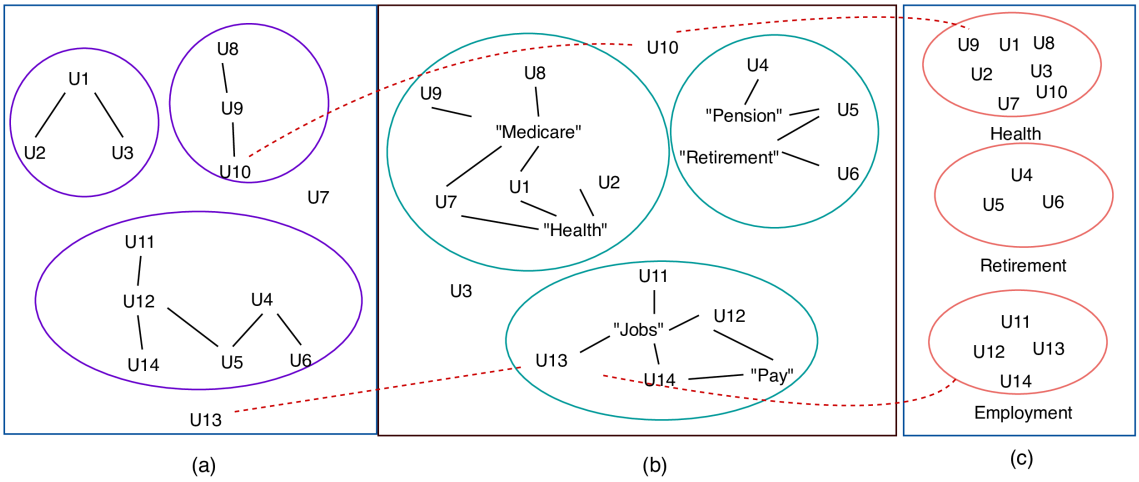


Figure 3.7: An example of (a) structure-based, (b) content-based, and (c) structure and content based communities

based on *what* content they share [54, 78] allows identification of user communities reflected in written media posts. However, these short texts are noisy and extremely sparse. Consequently, these content-based community detection methods produce inferior outcomes due to the curse of dimensionality [50].

In this paper, we conjecture that exploiting both the structure and content information effectively to identify user groups will achieve a meaningful solution. As shown in Fig. 3.7, due to the sparseness and noise, it is challenging to decide the communities of some users (e.g., $U13$ and $U3$) using a single type of input only. However, utilizing both structure and content together can produce more accurate results. A naive approach will be to concatenate all this information into a single feature matrix and apply clustering [87, 88]. Clustering on the concatenated data can embed all available information from all types into the learning process. However, this process usually results in poor outcomes as it ignores the inter-relatedness between sample objects and different feature objects, or between the sample objects, or between the feature object types [70, 99]. We present the community discovery problem as multi-type relational data (MTRD) learning, which simultaneously groups all data object types (e.g., users, tweets, etc.) considering all possible relationships existing between objects.

Community discovery in multiplex networks [76, 95] is one of the related research areas, which tries to identify communities using heterogeneous user relationships (e.g. friend, colleague). However, these methods do not take noisy and high-dimensional text data as input, which makes the community discovery process more challenging and complex. Some community discovery methods [6, 80, 84] discover groups of users by taking both content and structure into the learning process. However they require label information or impose many constraints. We propose a novel approach based on non-negative matrix factorization (NMF) in the unsupervised MTRD learning framework, named as Multi-Type Community Discovery (MTCD). We propose the MTRD learning with the added regularization to ensure that the global structure is preserved in the learned low-rank representations, i.e., two users who are similar (or dissimilar) in the original data should be kept close (or far) in the new low-dimensional space [72]. We empirically evaluate MTCD using three real-world Twitter datasets and benchmark with the state-of-the-art community discovery and clustering methods. The empirical analysis shows that MTCD can handle sparse datasets (both structural and content wise) and produce more accurate clusters than the benchmark methods. We also explore the different data representation and integration approaches and how they affect the outcome.

3.7 Related Work

Community discovery is a popular research problem which facilitates several real-world applications. In this paper, we propose to combine the structural relationship among users with the content similarity to generate meaningful user groups. With this objective, we categorize the related work as (1) Single-type and (2) Multi-type methods.

3.7.1 Single-type data methods

Researchers have predominately focused on exploiting graph models, derived from structural (user-user relationship) information, to form communities [17, 40, 86]. These methods are based on the theories behind network formation or system behavior which highlight *how* users are connected. However, in a sparse network, there can be many disconnected user groups interested in the same topic. Unfortunately, these structure-based methods identify these groups as distinct. Moreover, in a heterogeneous social network, user links are formed by different types of connections, e.g., being friends, relatives or colleagues, which lead to identifying user groups with mixed interests.

A very few researchers have analyzed the media content similarity to group common users [54, 78]. The traditional text analytic methods based on distance, density or probability face difficulties due to the noise and extreme sparseness of short text media data [75]. These methods result in poor outcome due to negligible distance differences in higher dimensionality [55] or information loss in lower dimensional projection [75].

In this paper, we propose to overcome these issues faced by using the structure or content data alone by combining these multiple types of data and finding the latent relationships among users.

3.7.2 Multi-type data methods

Multiplex network methods: In recent years, multiplex networks have been used to represent different relationships between the same set of users in a multi-layer network and various methods have been proposed for community discovery [76, 95]. Although these methods can handle the heterogeneity of networks, the sparseness issue remains unsolved. Moreover, multiplex networks are focused on representing different types of relationships among the same type of objects (ex. user-user).

They ignore the relationships among different types of objects (ex. user-hashtag, user-term) that is significant for grouping users, based on similar interest. Although it is possible to transform user-term relationships into a user-user relationship and derive multi-layer networks, the transformation process is time-consuming and looses some useful information in the process.

MTRD-clustering methods: Due to the easy availability of MTRD data such as in computer vision, bio-informatics and web-based system, a new area of multi-type relational data (MTRD) clustering has emerged [70, 73, 99]. These methods incorporate relationships between objects of different object types in the process. Object types can be data sample or different data features. MTRD methods based on non-negative matrix factorization (NMF) framework have been known to produce effective result when dealing with high dimensional and sparse data [72]. In this paper, we deem the community discovery problem as MTRD clustering and propose to simultaneously groups users by exploiting relatedness using both the content and structure information inherent in social media data. Prior researchers have used the NMF framework (but not in MTRD setting) to combine content and structure information [6, 80, 84]. Unfortunately, methods in [6, 84] require the community label information to learn the commonality amongst users. Method in [80] is unsupervised; however, it imposes many constraints as well as it requires many factor matrices to update. Imposing many constraints can help in achieving a meaningful solution, but it makes the learning process too complex and time consuming. It may also bring adverse effect to the learning process if coefficients are not set correctly. To the best of our knowledge, there exists no effective MTRD-based community mining method.

3.8 Multi-type Relational Data Learning for Community Discovery (MTCD)

3.8.1 MTCD: Problem Definition

Suppose $\mathcal{D} = \{X_1, X_2, \ldots, X_m\}$ be a MTRD dataset with m data types where X_h denotes the hth data type. Examples of data types in Twitter dataset can be retweet, reply, mention, tweet text, hash tags and URLs. Suppose $\{R_{hl} \in \mathbb{R}_+\}$ is a set of pairwise relationship matrices where $R_{hl} = \{r_{ij}\}^{n_h \times n_l}$, r_{ij} denotes the relationship (e.g., the *tf-idf* weight of a term in a tweet) between ith object and jth object of X_h and X_l respectively, $R_{lh}^T = R_{hl}$.

The task in this paper is to simultaneously group m object types into c clusters by considering the relationships between objects of different types in the dataset. The NMF-based objective function for MTRD can be written as [70, 99],

$$J = \min \sum_{1 \le h < l \le m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2, G_h \ge 0, G_l \ge 0 \tag{3.25}$$

where $\|.\|_F$ denotes the Frobenius norm and $G_h \in \mathbb{R}_+^{n_h \times c}$ and $G_l \in \mathbb{R}_+^{n_l \times c}$ are low-rank representations of object types X_h and X_l, respectively. Since G_h and G_l are being simultaneously optimized to be the low-rank representations of object types X_h and X_l, S_{hl} plays the role of a trade-off matrix that provides additional degree of freedom for the factorizing process [29].

The objective function in Eq. (3.25) will simultaneously learn the low-rank representations of sample object type and all feature object types. As reported in many high-order co-clustering methods [28, 57], the process of clustering of sample objects is benefited by using information of clusters of feature objects. Due to simultaneously learning the factor matrices of all feature object types and the sample object type during the learning process, the objective function creates an

effective learning process for learning the meaningful sample object clustering. The new low-rank representations are able to capture the cluster structures of all object types.

This objective function incorporates all inter-type relationships between users and other types such as the content they post. In social networking sites, there also exist many relationships between users, showing how they are connected. These relationships will return in symmetric input data matrices. Directly applying the objective function in Eq. (3.25), designed for an asymmetric matrix, will lead to impractical results. Therefore, we propose to add the second term to deal with the symmetric data matrices [61]. These two terms will learn the inter and intra multi-relationships present in the data respectively. However, prior research has shown that the NMF framework may need to include manifold learning to guide the learning process so that the global structure of the data remain preserved in the learned low-rank presentation [72].

In order to have more meaningful communities, we propose to implement the manifold learning by including an affinity matrix that will help in preserving the global structure for the learned low-rank representations, i.e., two users who are similar (or dissimilar) in the original data should be kept close (or far) in the new low-dimensional space. A third term is added in Eq. (3.25) to indicate this. Finally, we propose to add the pseudo-orthogonal constraints on factor matrices ($G_h^T D_h G_h = I$), which has shown to force the factor matrices to learn the optimal solution by learning the association among clusters of different object types [72]. The novel objective function for a community discovery problem can be defined as follows,

$$J = \min \sum_{1 \le h < l \le m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2 + \sum_{h=1..m, i=1..q} \|R_{hi} - G_h G_h^T\|_F^2 - \sum_{h=1}^{m} Tr(G_h^T W_h G_h),$$

$$\text{s.t., } G_h \ge 0, G_l \ge 0, S_{hl} \ge 0, G_h^T D_h G_h = I \tag{3.26}$$

Including the second term, the objective function learns that there exists q symmetric data matrices between members of object type h, corresponding to different q extents. Note that q can be zero when there exists no symmetric data matrix in the problem. The third term will ensure that all the important distances of data in the original space will be maintained in the new low-dimensional space. W_h is the affinity matrix built by using both the k nearest neighbour (kNN) and p farthest neighbour (pFN) graph with the purpose of encoding both close and far distance information [72], D_h is the diagonal matrix, $(D_h)_{ii} = \sum_j (W_h)_{ij}$.

3.8.2 Proposed Solution

We provide the solution for the objective function in Eq. (3.26) with respect to S_{hl}, G_h, G_l. We use the multiplicative update rule [64] where each variable will be separately updated while fixing others as constants until convergence [43, 66, 99].

1. Update rule for S_{hl}: When fixing h, l and thus fixing G_h, G_l, Eq. (3.26) is reduced to minimizing

$$J_{S_{hl}} = \|R_{hl} - G_h S_{hl} G_l^T\|_F^2,\ S_{hl} \geq 0 \tag{3.27}$$

By setting the derivative of $J_{S_{hl}}$ to be zero and using the condition for non-negative constraint [29], we have the update rule for S_{hl},

$$S_{hl} = (S_{hl})_{ij} \left[\frac{(G_h^T R_{hl} G_l)_{ij}}{(G_h^T G_h S_{hl} G_l^T G_l)_{ij}} \right]^{1/2} \tag{3.28}$$

2. Update rule for G_h, G_l: When fixing h, fixing $S_{hl}, G_l, h < l \leq m$, fixing $S_{lh}, G_l, 1 \leq l < h$, optimizing Eq. (3.26) with respect to G_h is equivalent to op-

timizing the following,

$$J_{G_h} = \sum_{h<l\leq m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2 + \sum_{1\leq l<h} \|R_{lh} - G_l S_{lh} G_h^T\|_F^2 + \sum_{1\leq i<q} \|R_{hi} - G_h G_h^T\|_F^2 - Tr(G_h^T W_h G_h) \quad (3.29)$$
$$G_h \geq 0, G_h^T D_h G_h = I,$$

With the pseudo-orthogonal constraint $G_h^T D_h G_h = I$, we can introduce the Lagrangian multiplier and take the first deviation on the Lagrangian function before using the Karush-Kuhn-Tucker (KKT) condition [18] for the non-negative constraint $G_h \geq 0$ to infer the update rule for G_h and similarly for G_l [72].

3.9 Empirical Analysis

We used several Twitter datasets obtained from TrISMA[5] focusing on Cancer, Health and Sports domains as reported in Table 3.5. We have chosen the set of groups under a domain where we can identify Twitter accounts to collect posts and interaction information. Each group in each domain is considered as the ground-truth community to benchmark the algorithmic outcome. For all the datasets, the user interaction networks are very sparse as shown by low-density values in Table 3.6. The conductance metric measures the fraction of total edge strength that points outside the ground truth communities. The conductance values of these datasets show the existence of different levels of inter-community interactions where sports and health datasets report the lowest and highest values respectively.

A clustering solution (or community) is evaluated by the standard pairwise F1-score which calculates the harmonic average of precision and recall, and Normalized Mutual Information (NMI) which measures the purity against the number of clusters

[5] https://trisma.org/

Table 3.5: Dataset Description

Dataset	#Users	#Interactions	#Tweets	#Terms	#Classes
DS1:Cancer	1585	1174	8260	2975	8
DS2:Health	2073	2191	19758	5444	6
DS3:Sports	5531	19699	12044	3558	6

Table 3.6: Density and Conductance of each Dataset

Dataset	Density	Conductance
DS1:Cancer	0.0005	0.152
DS2:Health	0.0005	0.274
DS3:Sports	0.0006	0.098

[119].

The empirical analysis was carried out with two objectives; (1) Identify an effective data representations for the proposed MTCD method and (2) Evaluate the effectiveness of MTCD by benchmarking the state-of-the-art community discovery and clustering methods.

3.9.1 Identify the effective data representation

In social media, two users can be similar based on their follower/friendship network or/and based on the people they interact and the content they share. The follower/friendship network is an outcome of heterogeneous relationships such as family, friend, colleague or fan. This network cannot be considered as a strong input for discovering communities based on common interests as it could yield communities with mixed interests. Moreover, extracting a complete follower/friendship network is a difficult task because of the data unavailability. Therefore in this paper, we propose to utilize the relationship between users with regard to their interactions as the structural input and the relationship between the user and media posts as the content input.

Interaction-Centric Relationships

In social media, users can interact with each other in many ways such as sharing others' posts, replying to other messages and mentioning others within their posts. We use three types of Twitter interactions and explore how these interactions should be represented as structural input in MTCD. For each pair of users, we count the total number of interactions between them and derive the relationship matrices R_{urt}, R_{ure}, R_{umt} and R_{us} which represent retweets, replies, mentions and combination respectively. The combined R_{us} matrix is created as follows,

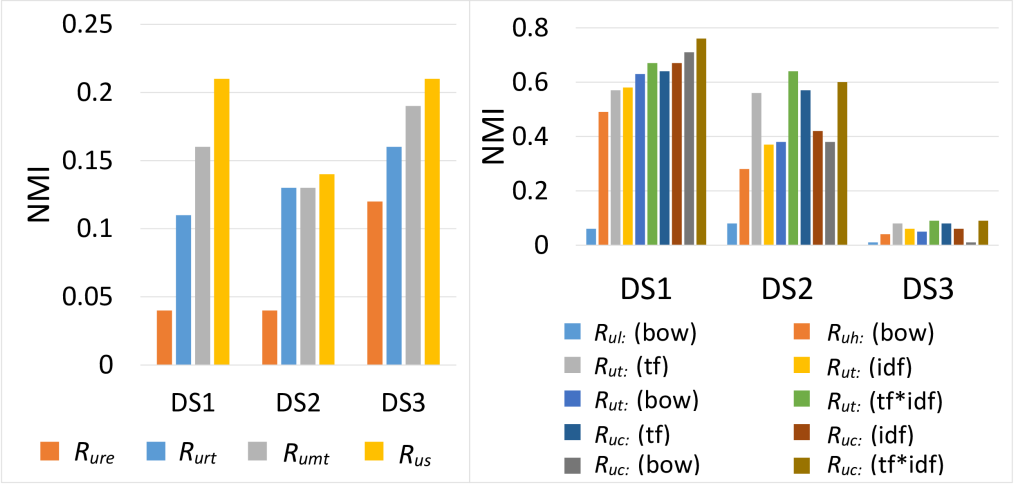
$$R_{us(ij)} = R_{urt(ij)} + R_{ure(ij)} + R_{umt(ij)} \tag{3.30}$$

where $R_{urt(ij)}$, $R_{ure(ij)}$ and $R_{umt(ij)}$ are the number of times users i and j retweet each others posts, reply to each other and mention each other respectively. R_{us} represents the relationships between objects of the same type, i.e., between users, covering the different types of interactions between users. R_{us} becomes the defined symmetric matrix R_{hi} in the proposed model in Eq. (3.26).

We test these different types of structural data to determine the effective structural representation in the NMF framework. Fig.3.8.a shows that the accuracy of NMF based clustering (measured as NMI) is highest with the combined interaction matrix (R_{us}). These results ascertain that all these interaction types indicate a similar community membership which makes the combination more effective. Therefore, we use the combined relationship matrix as the structural input to MTRD learning.

Content-Centric Relationships

We consider three kinds of Twitter content information for a user: text, hashtags, and URLs. Due to the post length restriction and high dimensionality, social media text are short in length and sparse in nature respectively [50]. Additionally, the external information (i.e., hashtags and URLs) in social media and unstruc-



(a) Performance based on NMF for different structure representations

(b) Performance based on NMF for different content representations

Figure 3.8: Performance based on NMF for different representations

tured phrases in text create new challenges in data representation for clustering purpose [50]. We use the standard pre-processing steps of stop words removal and lemmatizing for the text content. In order to deal with the unstructured nature and presence of abbreviations, we used a dictionary-based slang removal and word standardization.

The user-Hashtag R_{uh} and user-URL R_{ul} matrices are generated based on the count-based Bag-of-Words (bow) model (binary weighting scheme) as this information is rarely present. To represent the text content in media posts, the user-term matrix R_{ut} is generated with different term weighting schemes such as bow, term frequency *(tf)*, inverse document frequency *(idf)* and *(tf*idf)* [89]. To explore the effectiveness of combining content in a media post all the terms, urls and hash tags are considered as general terms and the user-content matrix R_{uc} is created as follows,

$$R_{uc(ij)} = w_{ij} \tag{3.31}$$

where w_{ij} represents the weight of i^{th} user having the j^{th} term. This value can

be represented using different weighting schemes such as bow, *tf*, *idf* and *tf*idf*. All these representations are tested in order to identify the best representation for MTCD. The R_{uc} matrix can be considered as the inter-relationship between user and content as denoted by R_{hl} in the proposed model in Eq. (3.26).

Fig. 3.8.b presents the accuracy of NMF based clustering (measured as NMI) with these different content representations. The *(tf*idf)* weighting scheme gives best NMI and makes the combined content representation R_{uc} represented with *(tf*idf)* as the best candidate to represent content relationship. This confirms the capacity of *(tf*idf)* in assigning high weight to discriminative words in a document.

3.9.2 Evaluate the effectiveness of MTCD

Benchmarks

We extensively benchmark MTCD with methods that can utilize both structure and content information, as well as methods that can utilize either structure or content only information. For the immediate comparison, we compare with the traditional MTRD-NMF [99] method as well as the state-of-the-art multiplex network methods GenLouvain [76] and PMM [95]. These two multiplex network methods however, only accept the symmetric user-user relationships as input. In order to represent content relationship (i.e., user-content matrix), which is not symmetric with the symmetric interaction relationship (i.e.,user-user matrix), we construct a user-user matrix from R_{uc}. This matrix R'_{uc} takes the number of overlapping terms between user i and user j using the respective vectors of users.

$$R'_{uc(ij)} = \sum (R_{uc(i)} \cap R_{uc(j)}) \tag{3.32}$$

Furthermore, we explore the ability to combine multiple relationships without explicitly using a multi-type method. The combined interaction relationship information (R_{us}) and content-based user-user matrix (R'_{uc}) return in a single concate-

nated relationship matrix defined as R_sc and the NMF framework is used to find common groupings.

$$R_{sc(ij)} = R_{us(ij)} + R'_{uc(ij)} \tag{3.33}$$

The state-of-the-art content-based clustering methods used are k-means [55], NMF [64] and LDA [16]. The state-of-the-art structure-based community discovery methods used are Louvain [17], InfoMap [86] and Clauset-Newman-Moore [40]. We also compare the NMF based method by using structure only information.

Table 3.7: Performance comparison of different community discovery methods

Method	**DS1**		**DS2**		**DS3**	
	NMI	**F1-score**	**NMI**	**F1-score**	**NMI**	**F1-score**
Multi type - Content and Structure						
MTCD # (R_{uc} and R_{us})	**0.79**	**0.79**	**0.76**	**0.83**	**0.31**	**0.41**
MTRD-NMF # (R_{uc} and R_{us})	0.67	0.69	0.37	0.45	0.27	0.39
NMF # (R_c)	0.34	0.42	0.27	0.41	0.03	0.33
GenLouvain (R'_{uc} and R_{us})	0.52	0.50	0.46	0.62	0.48	0.53
PMM (R'_{uc} and R_{us})	0.26	0.38	0.18	0.33	0.14	0.37
Single type - Content						
k-means # (R_{uc})	0.72	0.74	0.50	0.59	0.07	0.36
LDA* (R_{uc})	0.41	0.47	0.38	0.60	0.0	0.32
NMF# (R_{uc})	0.76	0.79	0.60	0.69	0.09	0.37
Single type - Structure						
Louvain (R'_{us})	0.32	0.4	0.24	0.4	0.44	0.49
InfoMap (R'_{us})	0.32	0.4	0.26	0.43	0.45	0.47
Clauset-Newman-Moore (R'_{us})	0.32	0.42	0.25	0.43	0.53	0.58
*Notes: # and * represent the methods execute with tf-idf and bow representations respectively*						

Accuracy Analysis

Results in Table 3.7 confirm that the proposed MTCD method is able to learn the most accurate and meaningful user communities. MTCD outperforms MTRD-NMF due to the inclusion of both the association relationship learning and the complete geometric structure learning. The performance gap is significant in DS2, where both content and structure representation is sparse, as shown in Table 3.6.

Table 3.8: Computational complexity of different community discovery methods

Methods	Complexity	Note
MTCD	O(vn^2)	*n: number of users* *d: number of features in a term matrix* *m: number of edges in the network* *k: number of communities* *v: number of types*
MTRD-NMF	O(vn^2)	
NMF	O(n^2)	
GenLouvain	O($nlogn$)	
PMM	O(n^3)	
k-means	O(n^{dk})	
LDA	O(nd^2)	
Louvain	O($nlogn$)	
InfoMap	O($nlog^2n$)	
Clauset	O(m)	

This confirms the suitability of MTCD in dealing with sparse multi-type data. It indicates that, for the heterogeneous dataset such as Twitter, the embedded regularizations used in MTCD with NMF help to obtain the improved outcome. It can be noted that both MTCD and MTRD-NMF methods perform substantially better than the conventional NMF (except NMF content only). Since MTRD methods discover user communities by considering the relationships of users with other feature object types, i.e., interaction and content, they achieve more accurate user groups. Though NMF on concatenated R_c matrix considers all related information in the learning process, it is unable to learn the latent features in both object types within the higher dimensional matrix. Thus, it fails to bring about a satisfactory result. Although the multiplex network based GenLouvain method outperforms single type-structure-based methods most of the time, it produces inferior results to both MTCD and MTRD-NMF methods, showing its ineffectiveness in handling sparse and high dimensional text data.

Within the content only methods, NMF which forms groups by factoring sparse matrix into lower dimension, performs better than the centroid-based partitional (k-means) and generative probabilistic (LDA) clustering methods. Due to the nature of social media, this (short text) data presents extreme sparseness and results in much higher dimensional term vectors with less term co-occurrence where distance calculations in partitional methods, as well as probability calculation in probabilistic

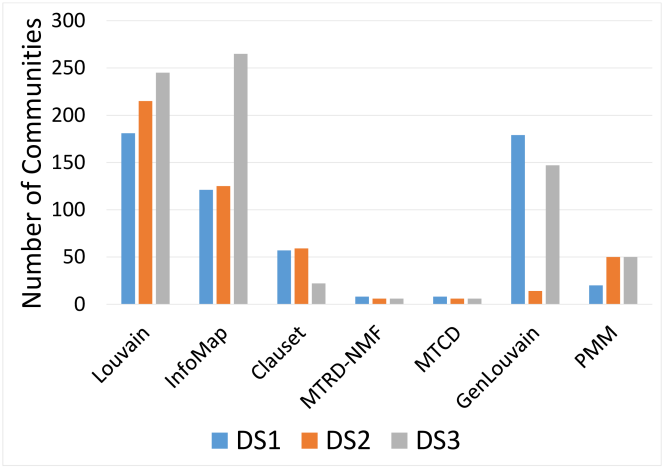


Figure 3.9: Number of communities detected in different community discovery methods

methods, are not able to succeed.

Structure-based methods are unable to outperform MTCD except in DS3 where the number of inter-community interactions is comparably high, as indicated by the conductance value. Mostly, MTCD outperforms single-type content and structure-based methods by complementing information given by each other. In DS3, as shown by the very poor performance of single-type content methods, they could not complement the structural information in a community discovery process and MTCD results in inferior performance than for structure-only methods.

Nevertheless, the structure-methods give a low-level picture of the network structure. As shown in Fig. 3.9, when the network is sparse, these methods generate a high number of communities, and they are not suitable for identifying high-level user communities with common interests. On the other hand, MTRD and other content-based methods were able to identify the smaller number of communities as per ground-truth. These content-based methods show better performance in identifying communities with comparably dense content representations. The proposed MTCD shows the capability of balancing these two, and is able to form meaningful communities in datasets where both structure and content are sparse.

Computational complexity

Results in complexity column of Table 3.7 show that all the NMF related methods have O(n^2) complexity or its' linear multiplication including MTCD. These are lesser complex than k-means and LDA clustering. In comparison, structure-based community discovery methods have less computational complexity. However, trade-off by achieving higher accuracy in terms of F1-score and NMI with MTCD as in Table 3.7 is well-justified for datasets with higher sparsity.

3.10 Conclusion and Future Work

In this paper, we present a novel approach of NMF-based MTRD learning to identify communities using both the structural and content information inherent in social media networks. This paper explores how the various types of information present on social media including media post as well as user interaction can be combined and included in the community discovery process. The proposed approach is evaluated on three Twitter datasets and benchmarked with several state-of-the-art methods. The experiment results show the importance of considering the community discovery problem in the heterogeneous context and learn it as multi-relational model, in order to achieve the accurate community groups and understand the behavior among user groups. The user-user relationship is considered as a special relationship in the proposed MTRD model. However, this important relationship should be paid more attention to obtain a more understanding. In the future, we will explore the applicability of MTRD learning for solving the overlapping community discovery problem.

Chapter 4

Learning the Diverse Manifold for Meaningful Multi-Aspect Data Representation

As illustrated in Figure 1.3 in Chapter 1, there are three different types of relationships in MTRD: inter-type relationship, intra-type relationship and association relationship. The previous chapter has devoted its attention to learning the accurate intra-type relationship and taking in the association relationship for MTRD data. This chapter will concentrate on the inter-type relationship with the target to learn more information from this relationship type under the perspective of how the inter-type relationship generates geometric shapes in data space.

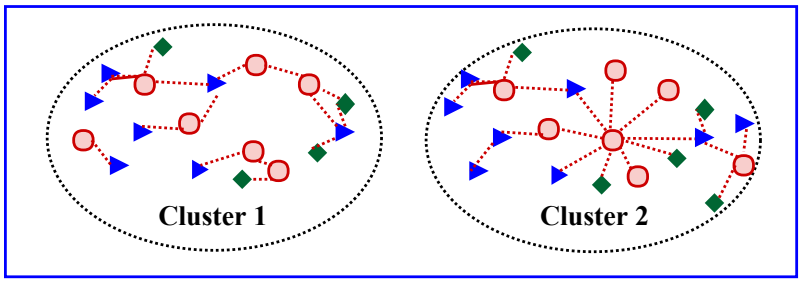


Figure 4.1: An example of a three type MTRD data with objects belonging to two clusters.

Figure 4.1 shows an example of MTRD data with three object types, e.g., documents, terms and concepts. Suppose there are two clusters in the dataset. Due to the high relatedness between data, the different type objects should belong to the two clusters as illustrated. By considering the geometric information created by the inter-type relationships between objects, together with the geometric shape created by the intra-type relationship, the data space will become a large graph that should be maintained when projecting to lower dimensional space.

This chapter presents the approach to learn and preserve the global geometric shape for multi-aspect data. This process is based on the NMF framework and known as learning the diverse manifold for multi-aspect data. The method considers the context of MTRD data where both manifolds generated from the intra-type relationship and inter-type relationship are incorporated into the learning process, in order to learn the diverse manifold and a meaningful representation for MTRD data. A new regularization term is added into the objective function to achieve this goal. The added term is also proved to embed the ranking term and help in selecting the most important features for the cluster learning process. The problem of learning MTRD is approached in a genetic manner, therein a comprehensive analysis of various important factors will be taken into consideration. The design of the method on multi-view data is also reported in this work, leading to an effective method for multi-aspect data clustering. This chapter is formed by Paper 4 in its original form.

Paper 4. Khanh Luong and Richi Nayak, Learning the Diverse Manifold for Meaningful Multi-Aspect Data Representation, IEEE Transactions on Knowledge and Data Engineering (TKDE) Journal (Under major revision).

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors:

Khanh Luong (PhD Candidate): Conceived the idea, designed and conducted experiments, analysed data, wrote the paper.

Signature: ***Date:*** 24-06-2019

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Signature: ***Date:*** 24-06-2019

ABSTRACT. Clustering on data with multiple aspects, such as multi-view or multi-type relational data, has become popular and useful in recent years due to the wide availability of this type of data. The approach using manifold learning with the Non-negative Matrix Factorization (NMF) framework, that learns the meaningful low-rank embedding of the multi-dimensional data, has shown effectiveness. We propose to incorporate the novel inter-manifold in the NMF framework, utilising the distance information of data points of different data types (or views) in order to learn the diverse and useful manifold for data learning. Empirical analysis reveals that the proposed method can find partial representations of various interrelated types and select useful features during clustering. Results on several datasets demonstrate that the proposed method outperforms the state-of-the-art multi-aspect data clustering methods in both accuracy and efficiency.

KEY TERMS. Multi-type Relational Data/ Clustering; Multi-view Data/ Clustering; Non-negative Matrix Factorization; Laplacian Regularization; Manifold Learning; Nearest neighbours;

4.1 Introduction

Multi-Aspect data, the data represented with multiple aspects, are becoming common and useful in practice [70]. This can be (1) *multi-view* data where samples are represented by multiple views; or (2) *multi-type relational data (MTRD)* where samples are represented by different data types and their inherent relationships. Aspects in MTRD are different object types (together with their associated relationships) and in multi-view data are multiple views. An example of MTRD and multi-view data are given in Figure 4.2.a and 4.2.b respectively. Figure 4.2.a. shows a MTRD dataset with 3 object types: Webpages, Terms and Hyperlinks with various relationships between objects of these object types. Figure 4.2.b. shows an example of two-view data where the first view is the representation of Webpages by Terms and

the second view is the representation of Webpages by Hyperlinks. The Webpages can be considered as samples while Terms and Hyperlinks are considered as features.

Though these two types of datasets might be treated similarly in most cases; there exists difference between them. The objective of MTRD data is to make use of as many relationships as possible to cluster all object types simultaneously, to enhance the performance of clustering of samples. In multi-view data, the relationship between different feature objects (e.g., between Terms and Hyperlinks) is ignored and the objective is to cluster samples using all view data and look for the consensus cluster structure. The multi-aspect data has shown to be more effective over the traditional single-type or single-view data, where each object is simply represented by a set of features [70]. An object type in MTRD is believed to provide different but valuable supplementary information for other types, due to their high relevance to each other[1] [120]. Inclusion of all object types and their correlations in MTRD clustering will provide a detailed view of data and will yield an accurate and meaningful clustering solution.

MTRD is represented by multiple types of objects and two main types of relationships, namely *inter-type* and *intra-type* [99]. Inter-type relationships contain connections between objects from different types (eg., co-occurrences between Webpages and Terms) and intra-type relationships contain the connections between objects of the same type (eg., similarities between Webpages). Each intra-type or inter-type relationship is normally encoded by a matrix. For instance, in Figure 4.2.a, three intra-type relationship matrices store intra-similarities between Webpages, between Terms and between Hyperlinks, and three inter-relationships matrices store relationships between Webpages and Terms, between Webpages and Hyperlinks and between Terms and Hyperlinks.

A naive way to make use of these intra-type and inter-type relationships is to

[1] In this paper, we will consider the context of MTRD and relate to multi-view data when necessary.

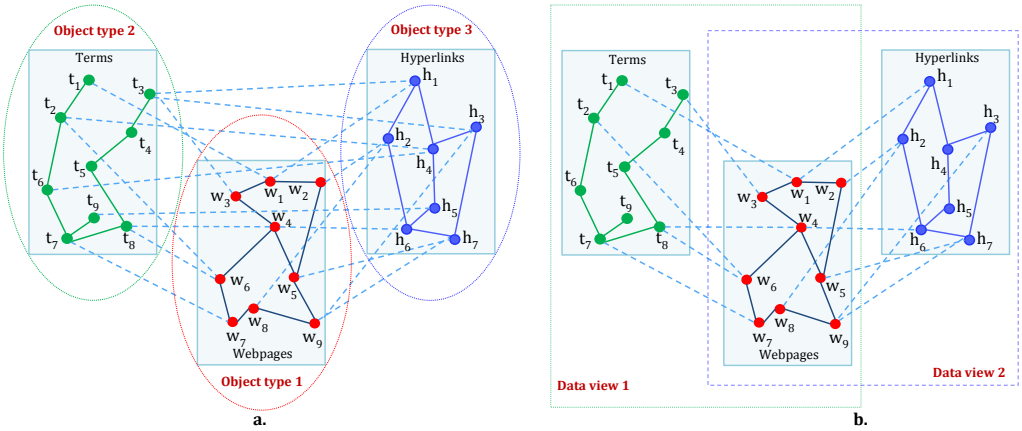


Figure 4.2: An example of Multi-Aspect Data. Fig. 4.2.a. A Multi-type Relational Data with three object types: Webpages, Terms, Hyperlinks. The intra-type relationships are represented as solid lines and inter-type relationships are represented as dotted lines. Fig. 4.2.b. A Multi-view Data example is represented with two views Terms and Hyperlinks

simply concatenate the matrices when applying a clustering algorithm. However, this approach loses the local structures explicitly embedded in intra- and inter-relationships and fails to utilize the complementary and compatible information [106]. Moreover, the real-world data is known to lie on multiple manifolds, which are embedded in the high and multi dimensional data space [10, 72, 120]. The approach of using manifold learning [11] with the NMF framework has shown a better outcome. The NMF framework is able to find the embedded low-dimensional space of original data where clusters can be found [21, 43, 66]. The combination of NMF and manifold learning is expected to learn the part-based representations that respect the embedded manifold of data, hence it results in meaningful clusters. Many efforts have been made to learn the accurate manifold for MTRD [48, 66, 72, 99, 100]. All these methods focus on learning the manifold on each object type only, i.e., considering how data points of the same object type reside on the manifold. Since MTRD exhibits multiple types of relationships, this leads to ignoring the intrinsic structure of the relatedness between data and learning the incomplete manifold. Due to the high relatedness between objects of different types, sampling of data points from different types should follow an intrinsic shape (or manifold). This should also

be learned and ensured for meaningful clustering. In this paper, we call this process *inter-manifold learning* to differentiate from *intra-manifold learning*, which focuses on the sampling of data points within the same data type on a manifold. To the best of our knowledge, no methods exist that exploit the residing on manifolds of data points from different types (e.g., the sampling on manifold of Webpages and Terms).

We propose to construct a p nearest neighbour (pNN) graph for each inter-relationship to capture close distances of points (high relatedness objects) from two different types, and aim to maintain these closenesses during the cluster learning process. The overarching aim is to meaningfully learn the diverse manifold generated on both intra-relationship and inter-relationship for original data and steadily preserve the learned complete manifold when mapping to the new low-dimensional data space. We propose a novel NMF framework that regularizes the complete and diverse geometric information that will play the role of an expert guiding the learning process. We design a well-defined and effective objective function for MTRD that enables learning the diverse manifold for multi-aspect data. We call this method of clustering the multi-aspect data as, learning Diverse Manifold for Multi-Aspect Data (DiMMA). Distinct from the existing methods [48, 66, 99, 100, 117], DiMMA is able to simultaneously find partial representations of various interrelated types, such that distance relationships between highly related objects from the same and different types are incorporated in the low-dimensional projected space. Empirical analysis with several datasets shows that DiMMA can cluster multi-aspect data accurately and efficiently. This is our first contribution.

Our second contribution lies in the ability of the proposed objective function of inter-manifold learning to identify and retain the most important features during the clustering process. This does not only help in significantly improving the clustering performance but also aids in dealing with high dimensionality, a problem inherent in multi-aspect data. We conducted an experiment to investigate the potential ability

of DiMMA to work on the high-dimensional multi-aspect dataset and it has shown to be effective.

In summary, we formulate and solve the problem of clustering MTRD by directly and simultaneously learning from the input data matrices, rather than relying on reformulating the big symmetric matrix [99] as for most of the other existing MTRD clustering methods [42, 48, 66]. For the newly formulated objective function, we offer a set of updated rules with guarantee on the correctness and convergence of the proposed algorithm.

The rest of the paper is organized as follows. Section 2 details the related work highlighting manifold learning and NMF-based methods for multi-aspect data. It also includes a review of the relationship between many vital and related concepts in clustering and manifold learning. Section 3 comprehensively presents the proposed DiMMA method. Section 4 presents an intensive empirical analysis. The conclusion and future work are in Section 5.

4.2 Related work

4.2.1 Manifold Learning

Manifold learning is long known for non-linear dimensionality reduction by preserving the local and/or global geometric structure of original data to the projected space. Some of the well-known methods are LLE [52] and Laplacian Eigenmaps [10] that focus on maintaining the local structure of data by ensuring the distances between points within a neighbourhood area. Distinctly, ISOMAP [97] focuses on maintaining the global structure of data by preserving not only the distances of close points but also distances of far points. The approach that combines NMF and manifold learning to obtain a meaningful low-dimension space and search clusters in the new space has been proved to be effective and has received much attention in recent years [21, 66, 99, 100, 117]. This family of methods uses a similarity graph to encode

the local geometry information of the original data and to discretely approximate the manifold so the sampling of data on manifold can be investigated [51].

To model the local neighbourhood relationships between data points, similarity graphs have been constructed by using two distance-based concepts, i.e., ϵ-neighbourhood and k-nearest neighbour [74]. By incorporating the nearest neighbour graph as an additional regularized term in NMF when projecting the data to lower embedded space, the close distances of data points are maintained during the learning process [11]. This corresponds to maintaining the high relatedness between points, in other words, maintaining the local structure/shape of the original data. In the manifold theorem, this corresponds to optimally learning and preserving the intrinsic manifold of the original data in a classification or clustering problem [10].

4.2.2 NMF Clustering and Manifold Learning

There are several reasons why NMF-based methods are popular, such as they can do both hard and soft clustering as well as the outputs of NMF being easily extended to various data types due to the nature of matrix factorization [3]. Early works that extend NMF [63] to two-type data [29] and to multi-type data [70] produce insufficient accuracy since they did not consider the local geometric structure, which is effective for identifying clusters. The later works using geometric information on two-type [42, 43, 67] and multi-type data [48, 72, 99] or multi-view [115, 120], achieved significant improvement.

It continues to be a non-trivial task to find exactly the embedding manifold; RHCHME [48] tries to learn accurate and complete intra-type relationships by taking into consideration the data lying on the manifold and belonging to subspace at once. In [120], by generating the intrinsic manifold of a multi-view data embedded from a convex hull of all the views' manifolds, the consensus manifold is learned via a linear combination of multiple manifolds of data from all views. Recently, ARASP [72] proposed a novel fashion of learning the MTRD manifold, where the close and

far distance information are embedded and preserved steadily for each data type. These works have proved the importance and necessity of manifold learning with clustering.

However, these above manifold-based methods mainly focus on using the geometric structure of each object type in MTRD or each view in multi-view data only. They fail to consider the sampling on manifolds of data points from different object types in MTRD or fail to consider the sampling on manifolds of data samples and data features in multi-view data. By ignoring the local geometric structure in multi-aspect data, these methods lead to poor accuracy. This paper attempts to exploit all geometric structure information from the original data graph and aims not only to save the structures of sub-graphs embedded in data points within each object type or view but structures of sub-graphs immersed in high-relatedness data points from different data types or views.

4.3 Learning Diverse Manifold for Multi-Aspect Data

4.3.1 Problem Definition

Let $D = \{X_1, X_2, ..., X_m\}$ be the dataset with m object types. Each object type $X_h = \{x_i\}_{(1 \leq i \leq n_h)}$ is a collection of n_h objects. Let $\{W_h\}_{1 \leq h \leq m}$be a set of intra-relationship matrices of m object types where $W_h = \{w_{ij}\}^{n_h \times n_h}$ is a weighted adjacency matrix resulted from building a k nearest neighbour graph (kNN graph) of object type X_h [21, 48, 99]. $\{R_{hl} \in \mathbb{R}_+\}_{1 \leq h \neq l \leq m}$ be a set of inter-relationship matrices where $R_{hl} = R_{lh}^T$, $R_{hl} = \{r_{ij}\}^{n_h \times n_l}$, r_{ij} denotes the inter-relationship between ith object of X_h and jth object of X_l (e.g., the *tf-idf* weight of a term in a document).

The task of clustering MTRD is to simultaneously group m object types into c clusters by using the inter-type relationships between objects of different types, i.e.,

$\{R_{hl}\}$ as well as considering the intra-type relationships between objects in the same type i.e., $\{W_h\}$. We assume that the objects in the dataset, which represent different types of features, tend to have the same number of underlying groups due to their high interrelatedness. This assumption allows us to set the same dimensionality for the newly mapped low-dimensional spaces of all object types.

The NMF-based objective function for MTRD [70, 99] can be written as,

$$J_1 = \min \sum_{1 \le h < l \le m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2, \; G_h \ge 0, G_l \ge 0 \tag{4.1}$$

where $\|.\|_F$ denotes the Frobenius norm, G_h and G_l are low-rank representations of object types X_h and X_l respectively, S_{hl} is a trade-off matrix that provides an additional degree of freedom for the factorizing process [29].

4.3.2 Learning the Inter Manifold

Methods using manifold learning in MTRD or multi-view data focus on learning the intra-manifold within each data type or data samples within each data view [48, 72, 120]. These methods preserve local geometric structure or close distances between data points within each data type or each view only. However, these methods fail to preserve the geometric structure created by the data points of multiple types or between multiple views. They leave data points of different types sampled on space without controlling their close distances and similarly, to the multi-view context. This leads to the disregard of relatedness information between data of different types or views, which is valuable for the challenging heterogeneous learning. We propose to learn and preserve the inter-manifold for MTRD, based on the assumption that *if two data points, belonging either to the same type or to different data types, are close in original data space (i.e., have high relatedness), their new representations in the new low-dimensional space must also be close.* We formulate the new objective function with the inter-manifold based on this assumption. Since the local geometric

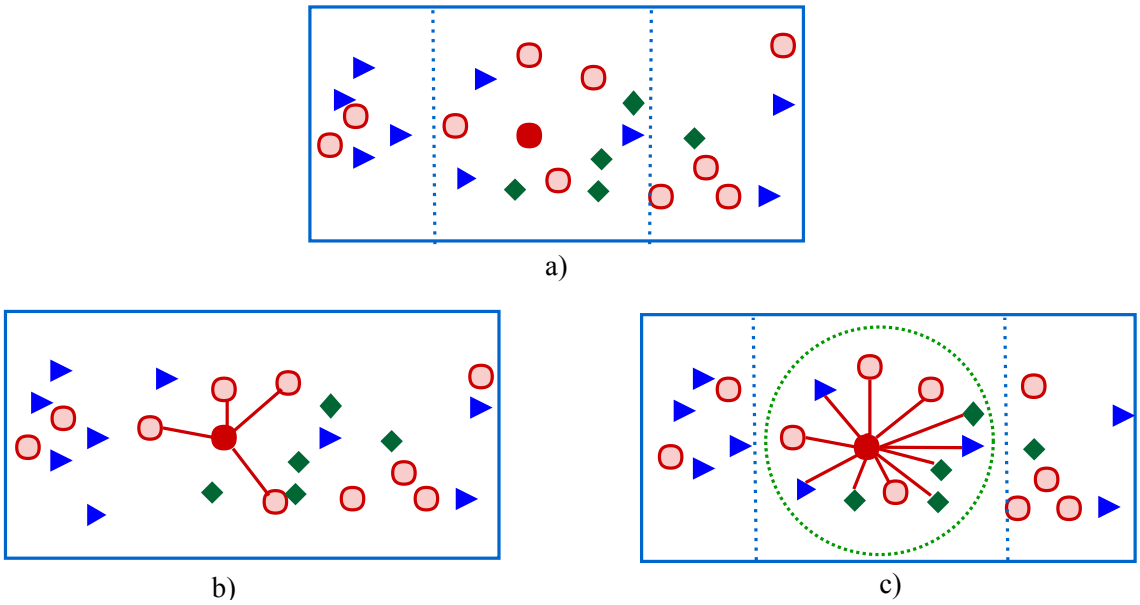


Figure 4.3: Illustration of how inter-manifold works on an MTRD dataset with three data types X_1, X_2 and X_3 lying on manifold in R^2. Figure 4.3.a shows the distribution of data in original space. Figure 4.3.b and Figure 4.3.c show the data distribution in the mapped low-dimensional space when only the intra-manifold has been used and when both intra- and inter-manifolds have been used, respectively.

information has been proved to be more useful to clustering task [10], we focus on building and maintaining the local geometric structure for inter-manifold only.

We propose to use the p nearest neighbour (pNN) graph for constructing a geometric shape of inter-manifold learned from the high dimensional original data. The closeness information between data points of two data types X_h and X_l is encoded in $Z_{hl} = \{z_{ij}\}^{n_h \times n_l}$ constructed as,

$$z_{ij} = \begin{cases} r_{ij} \text{ if } (x_j \in \mathcal{N}_p^{inter}(x_i) \text{ or } x_i \in \mathcal{N}_p^{inter}(x_j)) \\ 0, \text{ otherwise} \end{cases} \tag{4.2}$$

where $\mathcal{N}_p^{inter}(x_i)$ denotes p nearest neighbouring points from data type X_h of x_i.

Z_{hl} is the weighted adjacency matrix built from a pNN graph based on a *scatter of data points* of X_h and X_l. It stores *high inter-relatedness* between objects of X_h and X_l. In other words, constructing a pNN graph for each inter-type relationship matrix corresponds to the process of *discretely approximating* [43] the data points

of two object types on their intrinsic manifold. It learns distances from a point to all its neighbours from the other types.

To preserve this closeness information when mapping to a lower-order space, we propose to add the following term:

$$\min \sum_{h,l=1,h\neq l}^{m} \sum_{i=1}^{n_h} \sum_{j=1}^{n_l} \|g_i - g_j\|^2 z_{ij} \tag{4.3}$$

where $\|g_i - g_j\|^2$ is the Euclidean distance estimating closeness between new representations g_i, g_j projected from x_i, x_j, $x_i \in X_h$ and $x_j \in X_l$.

Suppose that h, l are fixed, Eq. (4.3) can be reduced to

$$\min \sum_{i=1}^{n_h} \sum_{j=1}^{n_l} \|g_i - g_j\|^2 z_{ij} \tag{4.4}$$

It is equivalent to,

$$\min\Big(\sum_{i=1}^{n_h} g_i g_i^T \sum_{j=1}^{n_l} z_{ij} + \sum_{j=1}^{n_l} g_j g_j^T \sum_{i=1}^{n_h} z_{ij} - 2 \sum_{i=1}^{n_h} \sum_{j=1}^{n_l} g_i g_j^T z_{ij} \Big) \tag{4.5}$$

$$\Leftrightarrow \min\big(Tr(G_h^T T_{hl}^r G_h) + Tr(G_l^T T_{hl}^c G_l) - 2Tr(G_h^T Z_{hl} G_l) \big) \tag{4.6}$$

where T_{hl}^r is a diagonal matrix of size $n_h \times n_h$ whose entries are sums of elements in each row of Z_{hl} and T_{hl}^c is a diagonal matrix size $n_l \times n_l$ whose entries are sums of elements in each column of Z_{hl}. More specifically,

$$(T_{hl}^r)_{ii} = \sum_{j=1}^{n_l} (z_{ij}), (T_{hl}^c)_{jj} = \sum_{i=1}^{n_h} (z_{ij}) \tag{4.7}$$

For all $1 \leq h \leq m, 1 \leq l \leq m, h \neq l$, and from equation (4.6), the optimization

problem in Eq. (4.3) becomes

$$\min \sum_{h,l=1,h\neq l}^{m} \big(Tr(G_h^T T_{hl}^r G_h) + Tr(G_l^T T_{hl}^c G_l) - 2Tr(G_h^T Z_{hl} G_l)\big) \tag{4.8}$$

$$\Leftrightarrow \min \big(\sum_{h=1}^{m} Tr(G_h^T T_h G_h) - 2 \sum_{1\leq h<l\leq m} Tr(G_h^T Q_{hl} G_l)\big) \tag{4.9}$$

where $Tr(.)$ denotes the trace of a matrix, T_h and Q_{hl} are defined as,

$$T_h = \sum_{l=1,l\neq h}^{m} (T_{hl}^r + T_{lh}^c) \tag{4.10}$$

$$Q_{hl} = Z_{hl} + Z_{lh}^T \tag{4.11}$$

Please refer to appendix A for full transformation from Eq. (4.8) to Eq. (4.9).

The newly proposed term (Eq. 4.9), we named P_1, is to learn *inter-manifold* and makes the proposed method distinct from other NMF-based methods [43, 48, 72, 99]. The new component P_1 ensures smoothness of the mapping function on objects from different types. Distances between objects (in the neighbourhood area) from different types are well preserved in the new mapped space. Figure 4.3 shows how the inter-manifold works to help in yielding a more meaningful representation of MTRD. While Figure 4.3.b illustrates the distribution of data when only the intra-manifold are considered, where highly related data points of different types have not been adequately respected. Figure 4.3.c illustrates when both intra- and inter-manifold learning are considered and the distances between a point to all its inter- and intra-neighbours have been preserved. It helps in learning true meaningful clusters.

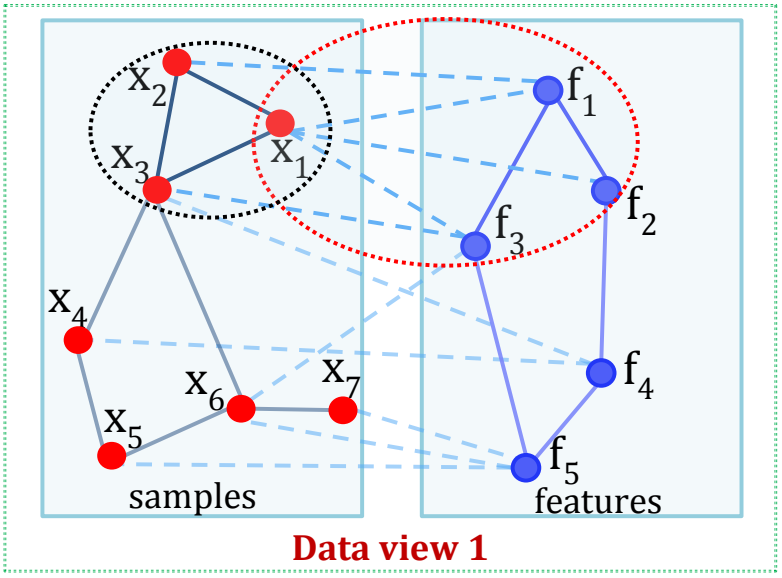


Figure 4.4: Illustration of learning the diverse manifold on multi-view data. Distances from a sample (x_1) to all its neighbouring samples (black dotted circle) and all its important features (red dotted circle) are maintained during the learning process.

When applied on the multi-view data, the proposed inter-manifold will play the role of a sample-feature manifold and will take care how sample objects and feature objects reside on manifolds to ensure the high relatedness between samples and features are preserved. Figure 4.4 illustrates a multi-view dataset, in which the sample-feature manifold is learned. Consider the data object x_1; by using sample-feature manifold, all important features of x_1 have been preserved and included during the learning process to the new low dimensional embedding space. This helps in obtaining the meaningful solution for multi-view data clustering.

The component P_1 can be integrated in the objective function 5.6. The combination of the inter-manifold term with the intra-manifold term generates a diverse manifold learning for MTRD.

The intra-manifold term on each object type in many recent methods [66, 99, 100] is given as,

$$P_2 = \min \sum_{h=1}^{m} (Tr(G_h^T L_h G_h)) \tag{4.12}$$

where the graph Laplacian matrix of object type X_h is defined as

$$L_h = D_h - W_h \tag{4.13}$$

D_h is the diagonal matrix computed by $(D_h)_{ii} = \sum_j (W_h)_{ij}$ and $W_h = \{w_{ij}\}^{n_h \times n_h}$ is a weighted adjacency matrix resulted from building a k nearest neighbour graph (kNN graph) of object type X_h [21, 48, 99]. Term P_2 (Eq. 4.12) ensures smoothness of mapping function on objects within the same type, and ensures that the distances between objects (in the neighbourhood area) from the same type are preserved in a new space.

We form the diverse manifold P encoding both inter-manifold Eq. (4.9) and intra-manifold Eq. (4.12) linearly as,

$$P = \delta P1 + \lambda P2 \tag{4.14}$$

The diverse manifold term P is incorporated into the objective function Eq. (5.6) to form the proposed and novel MTRD learning objective function as,

$$\begin{aligned} &\min \sum_{1 \le h < l \le m} \left(\|R_{hl} - G_h S_{hl} G_l^T\|_F^2 - \delta Tr(G_h^T Q_{hl} G_l) \right) \\ &+2\lambda \sum_{h=1}^{m} Tr(G_h^T L_h G_h) + \delta \sum_{h=1}^{m} Tr(G_h^T T_h G_h) \\ &\quad \text{s.t., } G_h \ge 0, G_l \ge 0, G_h 1_c = 1_{n_h}, G_l 1_c = 1_{n_l}, \forall 1 \le h < l \le m \end{aligned} \tag{4.15}$$

where $1_c, 1_{n_h}, 1_{n_l}$ are column vectors with sizes c, n_h, n_l, respectively, with all elements are 1s. The $l1$-normalization on G_h, G_l makes them more meaningful in ranking the values [48]. λ and δ are intra regularization and inter regularization parameters, respectively. If we set

$$Q_h = 2\lambda L_h + \delta T_h \tag{4.16}$$

Eq. (4.15) will become as follows:

$$J_2 = \min \sum_{1 \le h < l \le m} \left(\|R_{hl} - G_h S_{hl} G_l^T\|_F^2 - \delta Tr(G_h^T Q_{hl} G_l) \right) + \sum_{h=1}^{m} Tr(G_h^T Q_h G_h)$$
$$\text{s.t., } G_h \ge 0, G_l \ge 0, G_h 1_c = 1_{n_h}, G_l 1_c = 1_{n_l}, \forall 1 \le h < l \le m \tag{4.17}$$

The new objective function has three terms to simultaneously learn the low dimensional representations for all object types (the first term) that respect the intra and inter-manifolds (second and third terms respectively). Two parameters λ and δ can be adjusted. In particular, larger λ and δ values should be set if the dataset is known to be lying on manifolds. The m and $m(m-1)/2$ different values for λ and δ can be set to allow these parameters to act as weights of each object type and each inter-relationship type. In this paper, we give equal weight to each term to treat these object types and these relationships equally.

We summarize the important properties of the proposed diverse manifold in the following lemmas.

Lemma 1:

Minimizing the diverse manifold term P will help to ensure close distances (or high relatedness) between an object point to all its inter- and intra-neighbourhood points unchanged.

The definition of neighbourhood constraint is given as:

A pair of two points (x_i, x_j) is called to satisfy neighbourhood constraint if they satisfy inter- or intra-neighbourhood constraint. A pair of two points (x_i, x_j) belonging to two data types X_h, X_l is called to satisfy the inter-neighbourhood constraint if $x_j \in \mathcal{N}_{inter}^p(x_i)$ or $x_i \in \mathcal{N}_{inter}^p(x_j)$ where $\mathcal{N}_{inter}^p(x_i)$ denotes the p nearest inter-type neighbour points of x_i. A pair of two points (x_i, x_j) from the same data type X_h, is called to satisfy the intra-neighbourhood constraint if $x_i \in \mathcal{N}_{intra}^p(x_j)$ or

$x_j \in \mathcal{N}^p_{intra}(x_i)$, where $\mathcal{N}^p_{intra}(x_i)$ denotes the p nearest neighbour points of x_i.

Proof: We show that through minimizing P, the close distance between x_i and x_j will be preserved when x_i, x_j are being projected to lower dimensional space.

Considering an arbitrary pair of two points (x_i, x_j) that satisfies the neighbourhood constraint. We have:

(1) Firstly, the points (x_i, x_j) will show a high relatedness z_{ij}, or a close distance value $d(x_i, x_j)$ in the input space.

(2) Secondly, as (x_i, x_j) holds the neighbourhood constraint, from the definition of z_{ij} in Eq. (4.2) we have $z_{ij} > 0$. Minimizing $\|g_i - g_j\|^2 z_{ij}$, s.t. $z_{ij} > 0$ leads to minimizing $\|g_i - g_j\|^2$ or equivalently keeping distance $d(g_i, g_j)$ between g_i and g_j close.

From (1) and (2), we conclude that close distance between x_i, x_j in original space has been preserved on newly mapped points g_i, g_j in the new lower dimensional space through minimizing P. This fact is true on every pair of points that holds a neighbourhood constraint.

Most importantly, it should be noted that the value of z_{ij} is set to 0 when (x_i, x_j) is unable to satisfy the neighbourhood constraint, indicating that the points reside far away from each other. It ensures that the minimizing process only preserves the distances between data points that are close enough in the input space. It maintains the high relatedness between similar objects from the same type or from the different types while it ignores others.

Lemma 2:

Minimizing the diverse manifold term P will result in a ranking of features represented by objects of different types.

Proof:

As in equation (4.17), the objective function includes the following inter-manifold

term,

$$\begin{aligned} &\min \sum_{1\leq h<l\leq m} (-Tr(G_h^T Q_{hl} G_l)) \\ \Leftrightarrow &\max \sum_{1\leq h<l\leq m} Tr(G_h^T Q_{hl} G_l) \end{aligned} \tag{4.18}$$

Eq. (4.18) can be rewritten and represented as:

$$\sum_{1\leq h<l\leq m} Tr(G_h^T Q_{hl} G_l) = \sum_{h=1}^{m-1} \sum_{i=1}^{n_h} \sum_{l=h+1}^{m} q_i G_l g_i^{'} \tag{4.19}$$

where $g_i = [g_{i1}, g_{i2}, ..., g_{ic}]$ and $q_i = [q_{i1}, q_{i2}, ..., q_{in_l}]$

Consider a particular object $x_i \in X_h$, using Eq. (4.19), we have,

$$\sum_{l=h+1}^{m} q_i G_l g_i^{'} = \sum_{t=1}^{c} g_{it} \sum_{l=h+1}^{m} \sum_{j=1}^{n_l} q_{ij} g_{jt} \tag{4.20}$$

Let $Y = \sum_{l=h+1}^{m} \sum_{j=1}^{n_l} q_{ij} g_{jt}$. It can be seen that maximizing Eq. (4.20) will tend to assign a large g_{it} value to x_i if its related Y is large, due to the $l1$-norm constraint on G_h i.e., $\sum_{t=1}^{c} g_{it} = 1$. Furthermore, Y in Eq. (4.20) can be seen as a ranking term with regard to an object type. Considering features of objects of different types, assigning a higher value to g_{it} when Y has a promising high value will imply that the most important features were included during the clustering process.

This lemma shows that DiMMA supports feature selection while clustering and thus improving the quality and efficiency of the solution.

4.3.3 Algorithmic Solution to the DiMMA function

The DiMMA Algorithm

In this section, we provide the detail of the solution for the proposed objective function in Eq. (4.17) with respect to S_{hl}, G_h, G_l. We will separately update each variable while fixing others as constants until convergence [43, 66, 99] and introduce

the iterative algorithm 1 to simultaneously group several type objects into clusters.

1. Solving S_{hl}: When fixing h, l and thus fixing G_h, G_l, Eq. (4.17) with respect to S_{hl} is reduced to minimizing

$$J_{S_{hl}} = \|R_{hl} - G_h S_{hl} G_l^T\|_F^2 \tag{4.21}$$

$$\partial J_{S_{hl}}/\partial S_{hl} = -2R_{hl}G_h^T G_l + 2S_{hl}G_l^T G_l G_h^T G_h \tag{4.22}$$

$$\Leftrightarrow 2R_{hl}G_h^T G_l = 2S_{hl}G_l^T G_l G_h^T G_h \tag{4.23}$$

Then we have the update rule for S_{hl}

$$S_{hl} = (G_h^T G_h)^{-1} G_h^T R_{hl} G_l (G_l^T G_l)^{-1} \tag{4.24}$$

2. Solving G_h, G_l:

Solving $G_h, G_l, 1 \leq h < l \leq m$ is obviously equivalent to optimize all G_h where $1 \leq h \leq m$. When fixing h, fixing $S_{hl}, G_l, h < l \leq m$, fixing $S_{lh}, G_l, 1 \leq l < h$ we can rewrite the objective function in Eq. (4.17) as follows:

$$\begin{aligned} J_{G_h} &= Tr(G_h^T Q_h G_h) + \\ &\sum_{h<l\leq m} \left(\|R_{hl} - G_h S_{hl} G_l^T\|_F^2 - \delta Tr(G_h^T Q_{hl} G_l) \right) \\ &+ \sum_{1\leq l<h} \left(\|R_{lh} - G_l S_{lh} G_h^T\|_F^2 - \delta Tr(G_l^T Q_{lh} G_h) \right) \end{aligned} \tag{4.25}$$

By taking the derivative of J_{G_h} on G_h we have:

$$\begin{aligned}\partial J_{G_h}/\partial G_h &= 2G_hQ_h \\ &+ \sum_{h<l\leq m}\Big(-2R_{hl}G_lS_{hl}^T + 2G_hS_{hl}G_l^TG_lS_{hl}^T\Big) - \delta Q_{hl}G_l\Big) \\ &+ \sum_{1\leq l<h}\Big(-2R_{lh}^TG_lS_{lh} + 2G_hS_{lh}^TG_l^TG_lS_{lh} - \delta Q_{lh}^TG_l\Big)\end{aligned} \tag{4.26}$$

Let

$$\begin{aligned}A_h = &\sum_{h<l\leq m}(-R_{hl}G_lS_{hl}^T - 1/2\delta Q_{hl}G_l) \\ &+ \sum_{1\leq l<h}(-R_{lh}^TG_lS_{lh} - 1/2\delta Q_{lh}^TG_l)\end{aligned} \tag{4.27}$$

and

$$B_h = \sum_{h<l\leq m}(S_{hl}G_l^TG_lS_{hl}^T) + \sum_{1\leq l<h}(S_{lh}^TG_l^TG_lS_{lh}) \tag{4.28}$$

By introducing the Lagrangian multipler matrix Λ and setting the partial derivative of G_h to 0, we obtain

$$\Lambda = 2\lambda Q_hG_h - 2A_h + 2G_hB_h \tag{4.29}$$

Since the Karush-Kuhn-Tucker (KKT) condition [18] for the non-negative constraint on G_h gives $(\Lambda)_{ij}(G_h)_{ij} = 0$, we have the following update rule for G_h,

$$(G_h)_{ij} = (G_h)_{ij}\left[\frac{(Q_h^-G_h + A_h^+ + G_hB_h^-)_{ij}}{(Q_h^+G_h + A_h^- + G_hB_h^+)_{ij}}\right]^{1/2} \tag{4.30}$$

where $(Q_h^+)_{ij} = (|(Q_h)_{ij}| + (Q_h)_{ij})/2$, $(Q_h^-)_{ij} = (|(Q_h)_{ij}| - (Q_h)_{ij})/2$ and similarly to $A_h^{+/-}$, $B_h^{+/-}$ [30].

Algorithm 2 summarizes the proposed DiMMA approach. By iteratively updating G_h as in Eq. (4.30), DiMMA includes not only intra-type graphs but also includes inter-type graphs and returns the best result after convergence. The computational

steps are the same as in the NMF-based method incorporating intra-manifold learning, with the addition of the process of constructing $\{Z_{hl}\}_{1\leq h\neq l\leq m}$. Based on the concept of high relatedness or nearest neighbour, Z_{hl} can be built directly from R_{hl} without the need of computing distances. Consequently, this step significantly improves the overall complexity of the algorithm.

Convergence of the DiMMA Algorithm

We prove the convergence of the DiMMA Algorithm through update rules of S_{hl}, G_h as shown in Eqs. (4.24) and (4.30) respectively by using the auxiliary function approach [63].

Definition 2: *$Z(h, h')$ is an auxiliary function for $F(h)$ if the conditions $Z(h, h') \geq F(h), Z(h, h) = F(h)$, are satisfied* [63].

Lemma 3: *If Z is an auxiliary function for F, then F is non-increasing under the update* [63]

$$h^{(t+1)} = arg\min_{h} Z(h, h^{(t)})$$

Proof: $F(h^{(t+1)} \leq Z(h^{(t+1)}, h^{(t)}) \leq Z(h^{(t)}, h^{(t)}) = F(h^{(t)})$.

Lemma 4: *For any non-negative matrices $A \in \Re^{n\times n}, B \in \Re^{k\times k}, S \in \Re^{n\times k}, S' \in \Re^{n\times k}$, and A, B are symmetric, then the following inequality holds [30]*

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(AS'B)_{ip}S_{ip}^2}{S'_{ip}} \geq Tr(S^TASB)$$

Next, we will present two theorems that show the convergence of the DiMMA algorithm relating to the used auxiliary function.

Theorem 1: *Let*

$$L(G_h) = Tr(G_h^TQ_hG_h) - A_hG_h^T + G_hB_hG_h^T) \tag{4.31}$$

then the following function

$$
\begin{aligned}
&Z(G_h, G'_h) \\
&= \sum_{ij} \frac{(Q_h^+ G'_h)_{ij}(G_h)_{ij}^2}{(G'_h)_{ij}} - \sum_{ijk}(Q_h^-)_{jk}(G'_h)_{ji}(G'_h)_{ki}(1 + log\frac{(G_h)_{ji}(G_h)_{ki}}{(G'_h)_{ji}(G'_h)_{ki}}) \\
&-2\sum_{ij}(A_h^+)_{ij}(G'_h)_{ij}(1 + log\frac{(G_h)_{ij}}{(G'_h)_{ij}}) + 2\sum_{ij}(A_h^-)_{ij}\frac{(G_h^2)_{ij} + (G_h'^2)_{ij}}{2G'_{ij}} \\
&+\sum_{ij}\frac{(G'_h B_h^+)_{ij}(G_h^2)_{ij}}{(G'_h)_{ij}} - \sum_{ijk}(B_h^-)_{jk}(G'_h)_{ij}(G'_h)_{ik}(1 + log\frac{(G_h)_{ij}(G_h)_{ik}}{(G'_h)_{ij}(G'_h)_{ik}})
\end{aligned}
$$

is an auxiliary function for $L(G_h)$. It is a convex function in G_h and its global minimum is

$$(G_h)_{ij} = (G_h)_{ij}\left[\frac{(Q_h^- G_h + A_h^+ + G_h B_h^-)_{ij}}{(Q_h^+ G_h + A_h^- + G_h B_h^+)_{ij}}\right]^{1/2}$$

Proof: See Appendix B.

Theorem 2: *Updating $G_h, 1 \leq h \leq m$ using the update rule as in Eq. (4.30) will monotonically decrease the objective function as in Eq. (4.17)*

Proof: According to Lemma 3 and Theorem .1, we have

$$L(G_h^0) = Z(G_h^0, G_h^0) \geq Z(G_h^1, G_h^0) \geq L(G_h^1) \geq ...$$

Therefore $L(G_h)$ is monotonically decreasing.

Theorem 1 and theorem 2 guarantee the convergence of the DiMMA Algorithm regarding G_h. The sub problem in eq. (4.21) is a convex problem regarding to S_{hl} with the global minima in eq. (4.24). The correctness of the algorithm is also guaranteed as the update rules in Eq. (4.30) satisfies the Karush-Kuhn-Tucker optimal condition [18]. Thus, we conclude that DiMMA monotonically decreases the objective function in eq. (4.17) and converges to an optimal solution.

Algorithm 2: Learning Diverse Manifold for Multi-Aspect Data (DiMMA)

Input : Inter relationship matrices $\{R_{hl}\}_{1\leq h<l\leq m}^{n_h\times n_l}$, clusters number c, intra graph parameters λ, inter graph parameter δ, intra and inter-neighbourhood size k, p with non-negative values as discussed in the previous section.

Output: Cluster indicator matrices

Initialize non-negative matrices $\{G_h\}_{1\leq h\leq m}$ by K-means, using inter-relationship matrices $\{R_{hl}\}_{1\leq h<l\leq m}^{n_h\times n_l}$ and clusters number c as input.

1. for *each* $h, 1 \leq h \leq m$ **do**
construct Intra affinity matrices $\{W_h\}_{1\leq h\leq m}$ as in Eq. (3) [48]
compute the Laplacian matrices L_h as in Eq. (4.13).
end

2. for *each* $(h, l), 1 \leq h \neq l \leq m$ **do**
construct the affinity matrix Z_{hl} as in Eq. (4.2).
construct T_{hl}^r and T_{hl}^c as in Eq. (4.7)
end

3. for *each* $h, 1 \leq h \leq m$ **do**
construct T_h as in Eq. (4.10)
construct Q_h as in Eq. (4.16)
end

4. for *each* $(h, l), 1 \leq h < l \leq m$ **do**
construct Q_{hl} as in Eq. (4.11)
end

5. repeat
for *each* $(h, l), 1 \leq h < l \leq m$ **do**
update S_{hl} as in Eq. (4.24)
end
for *each* $h, 1 \leq h \leq m$ **do**
Update G_h as in Eq. (4.30)
Normalize G_h
end
until *converges*;

6. Transform $\{G_h\}_{1\leq h\leq m}$ into cluster indicator matrices by K-means.

Computational Complexity Analysis

Computational complexity of DiMMA includes three main components: Intra manifold learning; Inter manifold learning and Applying multiplicative updating.

The cost of learning intra-manifold is $O(n_h^2 km)$.

The cost of learning inter-manifold, steps $2-4$ in the DiMMA algorithm, is $O(n_h n_l m(m-1))$.

The cost of applying multiplicative updating includes the cost for updating $\{S_{hl}\}_{1\leq h<l\leq m}$, that is $O(n_h n_l c\frac{m(m-1)}{2})$ and updating $\{G_h\}_{1\leq h\leq m}$, that is $O(n_h^2 cm)$.

The overall time complexity of DiMMA can be expressed as:

$$O(n_h^2 km + n_h n_l m(m-1) + n_h n_l c\frac{m(m-1)}{2}))$$

Where m is the number of data types, k is the number of nearest neighbours, c is the number of clusters, and n_h, n_l are the number of objects of object types X_h, X_l, respectively.

The size of every object type is about the same. The values of m and c are comparatively much smaller than the object type size. Hence, the computational complexity of DiMMA remains quadratic. This is similar to the existing MTRD algorithms and the additional similarity computation will not incur extra cost, with the benefit of improving the accuracy. We show this in section 4 with extensive experiments.

4.4 Empirical analysis

4.4.1 Benchmarking methods

DiMMA is compared with the traditional NMF [64], DRCC co-clustering method [43] and its variation applicable to multi-type, and the state-of-the-art MTRD clus-

tering methods STNMF [99] and RHCHME [48], as well as the leading multi-view method MMNMF [117].

DRCC (Dual Regularized Co-Clustering) [43] is a co-clustering method that considers sampling on manifold for both data objects and features. It considers intra-manifold only. Its objective function is formulated as,

$$\begin{aligned} &\min \|R - GSF^T\|_F^2 + \lambda Tr(G^T L_G G) + \lambda Tr(F^T L_F F), \\ &s.t. G \geq 0, F \geq 0 \end{aligned} \tag{4.32}$$

where G and F are cluster assignment matrices for samples and features, S is a trade-off matrix that provides an additional degree of freedom for the factorizing process [29]. $L_G = D_G - W_G$ and $L_F = D_F - W_F$ are Laplacian graphs for samples and features, respectively.

To enable comparison between DRCC and DiMMA for clustering MTRD, we extend DRCC (known as **DRCC-Extended**) by constructing a graph Laplacian L_h for each object type X_h in the dataset. The revised objective function is defined as:

$$\begin{aligned} &\min \sum_{1 \leq h < l \leq m} \|R_{hl} - G_h S_{hl} G_l^T\|_F^2 + \lambda \sum_{1 \leq h \leq m} Tr(G_h^T L_h G_h) \\ &G_h \geq 0, G_l \geq 0, 1 \leq h < l \leq m \end{aligned} \tag{4.33}$$

$Tr(G_h^T L_h G_h)$ is the Laplacian regularization term that ensures smoothness of mapping function on object type X_h.

STNMF (Symmetric Non-negative Matrix Tri-Factorization) [99] is a leading MTRD clustering method and comes closest to DiMMA. It considers the intrinsic manifold on every object type by constructing a Laplacian graph corresponding to the data type. These Laplacian graphs of all data types are then combined into a new construction of Laplacian graph, as in equation (4.37). This newly formed Laplacian

graph is included in the proposed symmetric non-negative matrix factorization for processing clusters in the newly mapped space. Its objective function is defined as,

$$\min\|R - GSG^T\|_F^2 + \lambda Tr(G^T LG), s.t. G \geq 0 \tag{4.34}$$

where R is the symmetric inter-relationships matrix, G is the symmetric low-rank matrix and L is the newly constructed Laplacian graph matrix as defined below.

$$R = \begin{bmatrix} 0^{n_1 \times n_1} & R_{12}^{n_1 \times n_2} & \cdots & R_{1k}^{n_1 \times n_h} \\ R_{21}^{n_2 \times n_1} & 0 & \cdots & R_{2k}^{n_h \times n_2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{k1}^{n_h \times n_1} & R_{k2}^{n_h \times n_2} & \cdots & 0^{n_h \times n_h} \end{bmatrix} \tag{4.35}$$

$$G = \begin{bmatrix} G_1^{n_1 \times c} & 0^{n_1 \times c} & \cdots & 0^{n_1 \times c} \\ 0^{n_2 \times c} & G_2^{n_2 \times c} & \cdots & 0^{n_2 \times c} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_h \times c} & 0^{n_h \times c} & \cdots & G_h^{n_h \times c} \end{bmatrix} \tag{4.36}$$

$$L = \begin{bmatrix} L_1^{n_1 \times n_1} & 0^{n_1 \times n_2} & \cdots & 0^{n_1 \times n_h} \\ 0^{n_2 \times n_1} & L_2^{n_2 \times n_2} & \cdots & 0^{n_h \times n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0^{n_h \times n_1} & 0^{n_h \times n_2} & \cdots & L_k^{n_h \times n_h} \end{bmatrix} \tag{4.37}$$

The STNMF framework is effectively employed in RHCHME (Robust High-order Co-Clustering via Heterogeneous Manifold Ensemble) [48] when the objective function focuses on learning complete and accurate intra-relationship information by not only regarding data lying on manifolds but also concerning data belonging to subspaces. The RHCHME objective function proposed a new combination of Laplacian graph, i.e., $L = L^M + \lambda L^S$ where $L^M = D^M - W^M, L^S = D^S - W^S$, W^M and W^S are intra-relationship information learnt from the data sampling on manifolds and

subspaces, respectively.

These MTRD methods consider the intrinsic manifold of each object type in finding the low-dimensional representation and searching the clusters in the mapped space. Whereas, DiMMA not only preserves the intrinsic manifold on each object type (intra-manifold learning); it also preserves the intrinsic manifold embedded from the inter-relationships between objects of different data types (inter-manifold learning).

The multi-view method, MMNMF [117], learns the consensus coefficient matrix H_* by linearly combining the low-rank coefficient matrices of all views as:

$$J = \sum_{v=1}^{n_v} D_{KL}(X_v||H_v W_v) + \lambda Tr(H_*^T L_* H_*) \tag{4.38}$$

where $D_{KL}(.)$ is the Kullback-Leiber Divergence. $H_* = \sum_{v=1}^{n_v} \alpha_v H_v$. H_* is the consensus manifold learned by linearly combining the manifolds learned from all views as,

$$L_* = \sum_{v=1}^{n_v} \lambda_v L_v \tag{4.39}$$

For easy comparison with other NMF-based methods, we use Euclidean distance instead of Kullback-Leibler in Eq. (4.38). In addition to these MTRD and multi-view clustering methods, we compare DiMMA with the traditional NMF [64] to ascertain that MTRD methods give better performance in most cases, as they use more information.

4.4.2 Datasets

We use several real-world datasets to evaluate the performance of DiMMA (Table 4.1). Datasets MLR-1 (D1) and MLR-2 (D2) were created from the Reuters RCV1/ RCV2 Multilingual dataset [7] (or five view dataset). For the MTRD setting, object types include original English documents, English terms, French terms, German terms and Italian terms. Inter-type relationships are based on the ap-

pearances of each translated language terms in documents. Each intra-type relationship is generated based on the co-occurrence of corresponding terms in the same language documents. R-Top (D3) and R-MinMax (D4) were selected from the Reuters-21578, a well-known text dataset. D3 contains the 10 largest classes in Reuters-21578 and D4 includes 25 classes of the original dataset with at least 20, and at most, 200 documents per class. To create MTRD, we used external knowledge i.e., Wikipedia, and followed the process as in [48, 57] to generate the third data type concept (along with document and term data types). The processed dataset D3 presents two inter-relationships (or two views) in the form of documents-terms and documents-concepts. D4 uses three inter-relationships between documents and terms, documents and concepts; and between terms and concepts. The intra-affinity matrix on each object type created the following steps in [21, 43, 99]). Movie (D5) and RCV1-4Class (D6) datasets have been popularly used in clustering evaluation [15, 20, 53]. D5 contains the set of movies represented by two different views that are movies-actors and movies-keywords. D6, a subset of the RCV1 corpus [65], is a traditional dataset with no multi-type objects. D6 contains a large number of dimensions and is used to evaluate the performance of DiMMA against NMF in the presence of large data. NMF uses co-occurrences between documents and English terms on datasets D1 and D2; between documents and terms on D3, D4; and between movies and actors on D5. The co-clustering method DRCC uses the same information as NMF, but co-clusters both objects and features simultaneously.

4.4.3 Clustering results

We utilize the two widely applied measurement criteria, clustering accuracy (AC), the percentage of correctly obtained labels and Normalized Mutual Infor-mation (NMI) [119]. We also report the computational time of all methods. Average results are produced after 5-fold runs of each experiment. As presented in Tables 4.2-4.4, DiMMA outperforms all benchmarking methods on most datasets, except

D6 where RHCHME achieved higher performance. The reason behind DiMMA's superiority is its capability of utilising the local structures of objects from the same type and different types based on the proposed diverse manifold learning. This asserts the effectiveness of incorporating manifold learning on both intra and inter-type relationships into an NMF-based method. More importantly, on MTRDs D1-5, DiMMA performs exceptionally better as compared to the traditional clustering method NMF and co-clustering method DRCC. It justifies that if a method uses more information such as relationships with other object types effectively, it can achieve improved results.

Amongst the benchmarking methods, RHCHME achieves the second-best result on most datasets since it aims to learn the complete and accurate intra-type relationships. However, it is a complicated process of considering both data lying on manifolds and belonging to subspaces, thus the method often consumes much more time to converge compared to its counterparts (Table 4.4). MMNMF though, can produce good performance on some datasets (e.g., D1, D2), but it fails to match other MTRD methods on D4 since it fails to include the relationship between feature objects, i.e., between terms and concepts in this dataset. Both STNMF and DRCC-Extended exploit similar amounts of information, yet STNMF gives lower performance compared to DRCC-extended. This is because STNMF reformulates the inter-relationships information as a symmetric matrix instead of clustering directly and simultaneously on the inter-relationship matrices as DRCC that was extended to include the multiple types using the concept of DiMMA. We also observe that DRCC-extended always gives a better performance than DRCC, as the former uses more information in an efficient fashion. Moreover, on datasets D3-5, DRCC obtains higher performance compared to NMF as DRCC respects data and features structures during the clustering process. It is interesting to note that, on high-dimensional datasets D1, D2 and D6, the performance of DRCC, DRCC-Extended and STNMF is inferior to NMF, whereas DiMMA always outperforms

Table 4.1: Characteristic of the datasets

Properties	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
≠ Classes	6	6	10	25	17	4
≠ Object types	5	5	3	3	3	2
≠ Samples	8,400	3,600	4,000	1,413	617	9,625
≠ English terms	5,000	4,000	-	-	-	-
≠ French terms	5,000	4,000	-	-	-	-
≠ Germany terms	5,000	4,000	-	-	-	-
≠ Italian terms	5,000	4,000	-	-	-	-
≠ Terms	-	-	6,000	2,921	-	29,992
≠ Concepts	-	-	6,000	2,437	-	-
≠ Actors	-	-	-	-	1,398	-
≠ Keywords	-	-	-	-	1,878	-

Table 4.2: NMI of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
NMF	26.41	22.82	30.02	55.16	15.80	39.94	31.69
DRCC	15.01	14.14	46.04	63.45	22.41	31.75	32.13
STNMF	15.50	19.42	41.10	67.28	27.59	39.93	35.06
RHCHME	24.32	23.17	48.90	69.28	31.02	**50.46**	41.19
DRCC-Extended	22.43	20.60	46.27	71.64	29.57	31.75	37.04
MMNMF	30.86	30.59	44.11	60.50	23.11	-[1]	37.83
DiMMA	**34.96**	**30.76**	**50.38**	**73.06**	**33.87**	**46.59**	**44.94**

[1] Implementation of multi-view method MMNMF on one view dataset D6 is omitted.

Table 4.3: Accuracy of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
NMF	44.19	39.94	27.55	46.21	17.02	61.43	39.39
DRCC	31.80	24.58	49.85	53.22	22.53	48.94	38.49
STNMF	36.67	24.44	50.30	51.38	26.26	59.92	41.50
RHCHME	44.07	42.33	53.12	56.62	30.31	**72.98**	49.91
DRCC-Extended	43.19	39.75	47.25	59.45	26.90	48.94	44.25
MMNMF	45.23	44.19	**62.25**	45.65	20.58	-	43.58
DiMMA	**49.38**	**46.50**	51.10	**60.16**	**31.44**	**70.14**	**51.45**

Table 4.4: Running time (in thousand seconds) of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
NMF	0.79	0.26	1.06	0.05	0.015	7.54	1.62
DRCC	0.81	0.26	0.46	0.05	0.015	7.28	1.48
STNMF	9.46	5.00	1.63	0.26	0.060	8.33	4.12
RHCHME	127.59	36.83	1.63	0.26	0.065	8.22	29.10
DRCC-Extended	3.43	1.35	1.06	0.16	0.059	7.28	2.22
MMNMF	2.55	0.49	0.41	0.07	0.020	-	0.71
DiMMA	**2.55**	**1.39**	**1.06**	**0.16**	**0.057**	**7.75**	**2.16**

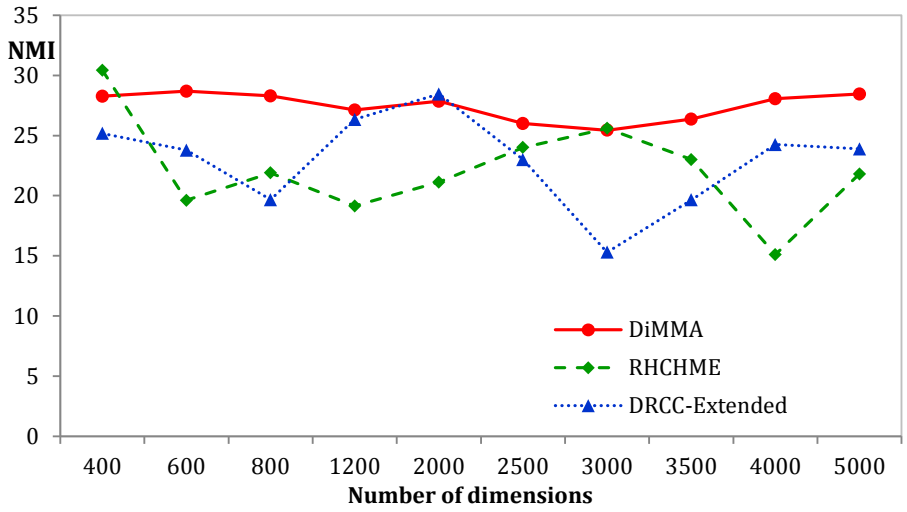


Figure 4.5: NMI curves of DiMMA and other MTRD clustering methods with the increase in dimensionality.

NMF. During clustering, DRCC, DRCC-Extended and STNMF incorporate intra-manifold learning that relies on a kNN graph, which may be challenging since the difference between distances from a particular point to its nearest and to its farthest points becomes non-existent in high-dimensional data, as noted by researchers previously [12]. DiMMA and RHCHME control this by including the inter-manifold learning or subspace learning, respectively.

With regard to time complexity, DiMMA consumes substantially less time than the other MTRD methods, STNMF and RHCHME. These methods are based on reformulating input matrices to be a symmetric matrix as in Eq. (4.35) thus require more running time, especially on large and high dimensional datasets such as D1, D2 and D6 (Table 4.4). Furthermore, as a consequence of the complex process to learn data subspace, RHCHME always consumes the most time. DiMMA consumes almost the same time as DRCC-Extended, but consumes more time than MMNMF and DRCC. However, the trade-off by obtaining a significantly improved accuracy justifies this.

We also observe that DRCC-extended always gives a better performance than DRCC, as the former uses more information in an efficient fashion. Moreover, on datasets D3-5, DRCC obtains higher performance compared to NMF as DRCC re-

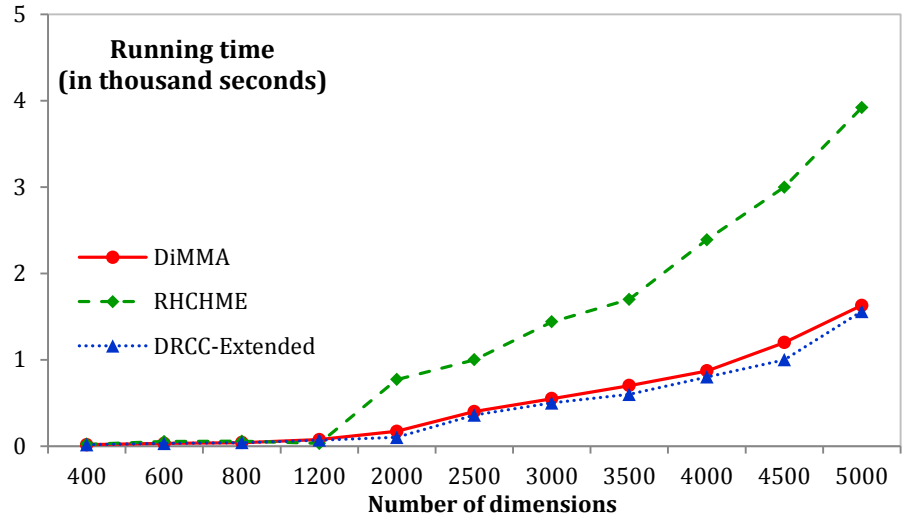


Figure 4.6: Performance time (in thousand seconds) of DiMMA and other MTRD clustering methods with the increase in dimensionality.

spects data and features structures during the clustering process. It is interesting to note that, on high-dimensional datasets D1, D2 and D6, the performance of DRCC, DRCC-Extended and STNMF is inferior to NMF. Whereas DiMMA always outperforms NMF. During clustering, DRCC, DRCC-Extended and STNMF incorporate intra-manifold learning that relies on a kNN graph, which may be challenging since the difference between distances from a particular point to its nearest and to its farthest points becomes non-existent in high-dimensional data, as noted by researchers previously [12]. Whereas DiMMA and RHCHME control this by including the inter-manifold learning or subspace learning, respectively.

With regard to time complexity, DiMMA consumes substantially less time than the other MTRD methods STNMF and RHCHME. These methods are based on reformulating input matrices to be a symmetric matrix as in Eq. (4.35), thus require more running time especially on large and high dimensional datasets such as D1, D2 and D6 (Table 4.4). Furthermore, as a consequence of the complex process to learn data subspace, RHCHME always consumes the most time. DiMMA consumes almost the same time as DRCC-Extended but consumes more time than MMNMF and DRCC. However, the trade-off by obtaining a significantly improved accuracy justifies it.

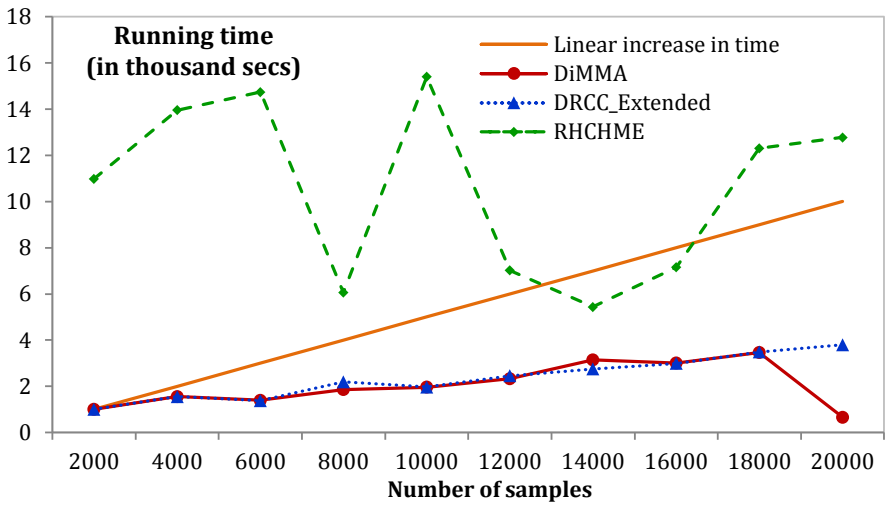


Figure 4.7: Scalability performance. The orange line shows the linear increase in time with the data size increase.

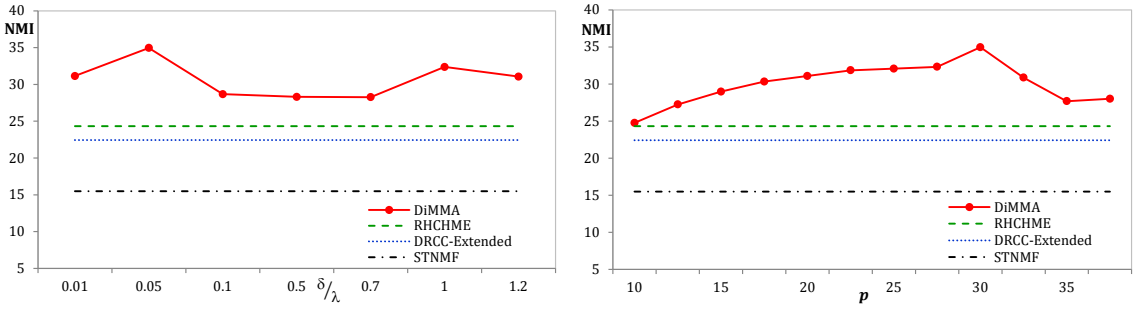


Figure 4.8: NMI curve with respect to inter-manifold learning regularization parameter δ and inter-neighbourhood size p on dataset MLR-1 (D1)

In summary, by considering the geometric structure of both intra-type and inter-type relations that guides the clustering process, the proposed DiMMA method is able to provide promising and meaningful clustering solutions.

4.4.4 Scalability of DiMMA

We investigate the impact of the size increase on the performances of DiMMA in comparison to other MTRD clustering methods. We select two benchmarking methods, DRCC-Extended and RHCHME, for this purpose. DRCC-Extended is the closest framework to DiMMA and RHCHME is the effective MTRD method based on the typical symmetric framework. We ignore NMF and DRCC in this experimental investigation since they are single or bi-type based methods, STNMF should

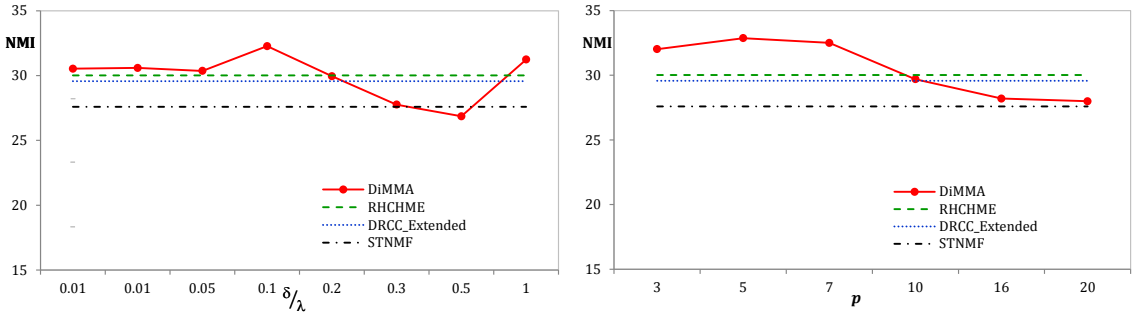


Figure 4.9: NMI curve with respect to inter-manifold learning regularization parameter δ and inter-neighbourhood size p on dataset Movie (D5)

have the same behaviour as RHCHME and MMNMF is a multi-view based method. We conducted experiments on a series of datasets that are extracted from Reuters RCV1/RCV2 Multilingual for this investigation.

To check dimensionality effects, a dataset with 600 samples and 400 features on each view, i.e., English term, French term, German term and Italian term, was chosen as a starting point and then feature dimensions were increased systematically up to 5000. As can be seen from Figures 4.5, 4.6, DiMMA requires much less time than RHCHME and about the same time as DRCC-Extended to produce a high and stable result over the datasets with the increase in dimensionality. This is due to the inclusion of inter-manifold learning where it constructs a pNN graph Z as in equation (4.2) without needing to calculate similarity but using the available inter-relationship information as a distance approximation.

To check the effect of collection size, another dataset with 2000 English documents was selected and the size was multiplied (Figure 4.7). The numbers of English, French, German and Italian terms are kept fixed at 5000. As shown in Figure 4.7, DiMMA performs better than the linear time increase. DRCC-Extended also performs similarly, however, RHCHME consumes much longer than the linear time increase.

4.4.5 Parameters setting

To enable a fair comparison between methods, we set the same values for all common parameters among methods. The intra-type neighbourhood size k for constructing a kNN graph for each object type is fixed to 5, intra-type regularization parameter λ is set to 10 for datasets D1, D3 and D5, and λ is set to 1 for other datasets. Apart from λ and k, DiMMA has two parameters, the inter-regularization parameter δ and neighbourhood size p. We will discuss how to choose these parameters to obtain the highest performance. Since DiMMA aims at incorporating both intra- and inter-manifold learnings to preserve local structures of data points from the same as well as different types, the parameters for intra- and inter-manifold can be set the same, i.e., we treat intra- and inter-manifold learning equally. However, the best results can be achieved through setting a right value for δ depending on real situations. As per our experiment, δ can be picked varied such that the ratio δ/λ is from 0.01 to around 1 depending on datasets.

With regard to choosing the neighbourhood size p, in the same way as other methods that rely on neighbourhood concept, good results can be obtained with a right neighbourhood size setting. In our experiments, the suitable p value can be set from 5 to 30. Figures 4.8- 4.9 show how the NMI values of DiMMA change with altering δ and p on datasets D1 and D5. Results validate that DiMMA produces stable results and easily achieves the highest values when the right values have been chosen for these parameters. We note that the parameters δ and p are not applicable for benchmark methods, hence there is no change with the alterations of these parameters while DiMMA curves stay above other benchmarks in most cases, as shown on these figures.

4.5 Conclusion

We presented a novel method DiMMA that provides a new look at the relation between two types of multi-aspect data, MTRD and multi-view, for the problem of clustering based on NMF framework along with a set of update rules, which lead to correctness and convergence. DiMMA incorporates the diverse geometric structures of intra-type and inter-type relationships in MTRD and the geometric structures of samples and samples-feature in multi-view data. Using the novel proposed diverse and complete manifold, this paper approaches the problem of multi-aspect data clustering in the most generic way, and provides a comprehensive analysis of vital factors that affect the problem, such as ranking terms, relatedness between objects, data points' distances and geometric structures.

Experimental results on many datasets with different sizes and different numbers of data types show the effectiveness of DiMMA over the relevant state-of-the-art methods. DiMMA shows a superior performance on both types of multi-aspect datasets. While most multi-view methods fail to work on MTRD, since they cannot capture the relationships between feature objects, MTRD-designed methods can work well on both types of data.

In future work, we will decompose the affinity matrices and form an overall framework for NMF clustering with graph regularization. Moreover, in the paper we have mentioned three terms to improve performance of a clustering method, i.e., intra-manifold learning, inter-manifold learning and subspace learning; we will investigate how the characteristic of a dataset relates to choosing the right parameters for the three terms in order to achieve the highest performance for clustering.

4.6 Appendix A: Arriving to Eq. (4.9) from Eq. (4.8)

.

Decomposing the last term from Eq. (4.8) into two parts, we have the following equivalent equation,

$$\begin{aligned}\min \big(& \sum_{h,l=1,h\neq l}^{m} Tr(G_h^T T_{hl}^r G_h) + \sum_{h,l=1,h\neq l}^{m} Tr(G_l^T T_{hl}^c G_l) \\ & -2 \sum_{1\leq h<l\leq m} Tr(G_h^T Z_{hl} G_l) - 2 \sum_{1\leq l<h\leq m} Tr(G_h^T Z_{hl} G_l)\big)\end{aligned} \tag{4.40}$$

Changing the roles of h and l in terms 2 and 4 leads to the following equation:

$$\begin{aligned}\min \big(& \sum_{h,l=1,h\neq l}^{m} Tr(G_h^T T_{hl}^r G_h) + \sum_{h,l=1,h\neq l}^{m} Tr(G_h^T T_{lh}^c G_h) \\ & - 2 \sum_{1\leq h<l\leq m} Tr(G_h^T Z_{hl} G_l) - 2 \sum_{1\leq h<l\leq m} Tr(G_l^T Z_{lh} G_h)\big)\end{aligned} \tag{4.41}$$

$$\begin{aligned}\Leftrightarrow \min \big(& \sum_{h,l=1,h\neq l}^{m} Tr(G_h^T T_{hl}^r G_h) + \sum_{h,l=1,h\neq l}^{m} Tr(G_h^T T_{lh}^c G_h) \\ & - 2 \sum_{1\leq h<l\leq m} Tr(G_h^T Z_{hl} G_l) - 2 \sum_{1\leq h<l\leq m} Tr(G_h^T Z_{lh}^T G_l)\big)\end{aligned} \tag{4.42}$$

Eq. (4.41) can be written as Eq. (4.42) due to the property of $Trace$, $Tr(G_l^T Z_{lh} G_h) = Tr(G_h^T Z_{lh}^T G_l)$.

Finally, we have Eq. (4.9) by replacing T_h and Q_{hl} as defined in Eq. (4.10), (4.11) into Eq. (4.42).

4.7 Appendix B: Proof of Theorem 1

To simplify the notations, we drop index h in the notation and re-write Eq. (4.31) as follows,

$$\begin{aligned} L(G) = Tr(&G^T Q^+ G - G^T Q^- G - G^T A^+ \\ &+ G^T A^- + G B^+ G^T - G B^- G^T) \end{aligned} \tag{4.43}$$

To prove Theorem 1, we need to first prove $Z(G, G')$ is an auxiliary function of $L(G)$ and then prove that $Z(G, G')$ converges and its global minimum is the update rule of G.

1. Prove $Z(G, G')$ is an auxiliary function of $L(G)$

Using Lemma 4 we have,

$$Tr(G^T Q^+ G) \leq \sum_{ij} \frac{(Q^+ G')_{ij} G_{ij}^2}{G'_{ij}} \tag{4.44}$$

$$Tr(G B^+ G^T) \leq \sum_{ij} \frac{(G' B^+)_{ij} G_{ij}^2}{G'_{ij}} \tag{4.45}$$

In addition, we have $a \leq \frac{a^2 + b^2}{2b}, \forall a, b > 0$, thus

$$Tr(G^T A^-) = \sum_{ij} A_{ij}^- G_{ij} \leq \sum_{ij} A_{ij}^- \frac{G_{ij}^2 + G_{ij}'^2}{2G'_{ij}} \tag{4.46}$$

We also have the inequality $z \geq 1 + \log z, \forall z > 0$. Therefore we have following inequalities

$$Tr(G^T A^+) \geq \sum_{ij} A_{ij}^+ G'_{ij} (1 + \log \frac{G_{ij}}{G'_{ij}}) \tag{4.47}$$

$$Tr(G^T Q^- G) \geq \sum_{ijk} (Q^-)_{jk} G'_{ji} G'_{ki} (1 + \log \frac{G_{ji} G_{ki}}{G'_{ji} G'_{ki}}) \tag{4.48}$$

$$Tr(GB^{-}G^{T}) \geq \sum_{ijk} B^{-}_{jk} G'_{ij} G'_{ik} (1 + \log \frac{G_{ij} G_{ik}}{G'_{ij} G'_{ik}}) \tag{4.49}$$

From Eqs. (4.44-4.49) we have $L(G) \leq Z(G, G')$. Furthermore we have $L(G) = Z(G, G)$ is obviously.

Thus $Z(G, G')$ can be called as an auxiliary function of $L(G)$ (Definition 2).

2. Prove $Z(G, G')$ converges and its global minimum is the update rule of G

We use Hessian matrix, and need to prove that the Hessian matrix of $Z(G, G')$ is a positive definite diagonal matrix.

Taking the first derivative of $Z(G, G')$ on G, we have

$$\begin{aligned} \frac{\partial Z(G, G')}{\partial G_{ij}} =& 2\lambda \frac{(Q^{+}G')_{ij} G_{ij}}{G'_{ij}} - 2\lambda (Q^{-}G')_{ij} \frac{G'_{ij}}{G_{ij}} \\ & - 2A^{+}_{ij} \frac{G'_{ij}}{G_{ij}} + 2A^{-}_{ij} \frac{G_{ij}}{G'_{ij}} \\ & + 2\frac{(G'B^{+})_{ij} G_{ij}}{G'_{ij}} - 2(G'B^{-})_{ij} \frac{G'_{ij}}{G_{ij}} \end{aligned} \tag{4.50}$$

The Hessian matrix of $Z(G, G')$ containing the second derivatives as in eq. (4.51) is a diagonal matrix with positive entries.

$$\begin{aligned} \frac{\partial^2 Z(G, G')}{\partial G_{ij} \partial G_{kl}} =& \delta_{ik}\delta_{jl} (2\lambda \frac{(Q^{+}G')_{ij}}{G'_{ij}} + 2\lambda (Q^{-}G')_{ij} \frac{G'_{ij}}{G^2_{ij}} \\ & + 2A^{+}_{ij} \frac{G'_{ij}}{G^2_{ij}} + 2\frac{A^{-}_{ij}}{G'_{ij}} \\ & + 2\frac{(G'B^{+})_{ij}}{G'_{ij}} + 2(G'B^{-})_{ij} \frac{G'_{ij}}{F^2_{ij}}) \end{aligned} \tag{4.51}$$

Thus $Z(G, G')$ is a convex function of G. Therefore, we can obtain the global minimum by setting $\partial Z(G, G')/\partial G_{ij} = 0$. After some transformations and replacing index h, we can get eq. (4.30).

Chapter 5

Learning Consensus and Complementary Information for Multi-view Data Representation

The problem of multi-view data learning is an emerging problem, due to the amount of the dataset type that has been arisen, and its advantages of carrying more valuable information, as compared to traditional one-view data. Due to the property of being represented by multiple views, the multi-view data is known to be very high-dimensional and sparse. A wide range of methods developed working on this data are based on factorization technique, i.e., NMF, to learn the low-rank representation before conducting the appropriate learning task from the data. To be applied on multi-view data, the NMF is first carried out on all available data views and the integration to learn the consensus embedding conducted in the next step. The approach has been the most effective approach for high-dimensional and sparse multi-view data for a long time, especially when it is incorporated with the manifold learning technique to learn and maintain the intrinsic geometric shape for data when learning the low-rank order. However, the challenge of applying NMF on multi-view

data lies at the step of integrating data to seek the consensus data representation from all views. Due to being learned independently from all views, different low-rank representations from each view have embedded latent features of data on their corresponding view only. This may lead to the loss of the interrelatedness among all views and will lack success to maintain the intrinsic consistent and complementing valuable information in multi-view data.

This chapter will present two proposed multi-view learning methods, MVCMF (Multi-view Coupled Matrix Factorization) and 2CMV (Learning Consensus and Complementary Information for Multi-view Data) in their original forms of two papers (Paper 5 and Paper 6), which concentrate on learning the advantageous low-rank representation for multi-view data.

Paper 5. **Khanh Luong** and Thirunavukarasu Balasubramaniam and Richi Nayak (2018), *A novel technique of using Coupled Matrix Factorization and Greedy Coordinate Descent for Multi-view Data Representation.* Lecture Notes in Computer Science: The 19th International Conference on Web Information Systems Engineering – WISE 2018.

Paper 6. **Khanh Luong** and Richi Nayak, *Learning the Consensus and Complementary Information for Multi-View Data Clustering.* In 21st International Conference on Computer Vision - ICCV 2019 (Under Review).

While Paper 5 focuses on learning the accurate consensus low-rank matrix for multi-view data with the aid of Coupled Matrix Factorization (CMF), Paper 6 focuses on learning the consensus low-rank matrix such that the complementary information will also be enhanced during the learning process. More specifically,

Paper 5 (containing detail on MVCMF) is presented in the first part of the chapter. MVCMF utilizes a CMF model and therefore allows the learning process to naturally and directly learn the common consensus latent features from multi-view data and ensure the learned low-rank representation embeds the compatible information for multi-view data. This is also the first work where the affinities

between data samples are also integrated in coupled factorizing in the form of a new addition view. Distinct from most existing factorization-based clustering methods, the MVCMF method offers to use the greedy coordinate descent scheme in the optimizing process, for the purpose of favouring the learning process with the aid in selecting the most important elements to be updated.

Paper 6, which proposes 2CMV, is presented in the second part of the chapter. It is inspired by paper 5, which has proposed a uniformed framework to naturally and effectively learn the consensus low-rank matrix for multi-view data using CMF, yet neglects to take care of the complementary information for the learning process. 2CMV is proposed as an integrated framework using both the NMF and CMF models to learn both consensus and complementary components for multi-view data. A novel complementary constraint is also proposed and imposed on the two components to allow the learning process to learn the most compatible and complementary information. More importantly, a novel optimal manifold concept for multi-view data is also introduced in 2CMV, leading to a well-defined factorization-based method for effective multi-view learning.

Next, the chapter will present two papers. Since this is a thesis by publication, each paper will be presented in its original form; due to their different formats, there will be some minor format differences. However, these do not change the content of the original papers.

Paper 5. Khanh Luong and Thirunavukarasu Balasubramaniam and Richi Nayak (2018), *A novel technique of using Coupled Matrix Factorization and Greedy Coordinate Descent for Multi-view Data Representation.* The 19th International Conference on Web Information Systems Engineering – WISE 2018.

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors

Khanh Luong (PhD Candidate): Conceived the idea, designed and conducted experiments, analysed data, wrote the paper and addressed reviewers comments (for published papers) to improve the quality of paper.

Signature: ***Date:*** 24-06-2019

Thirunavukarasu Balasubramaniam (PhD Candidate): Developed a module of the method, wrote a sub section in the manuscript and carried out experiment for the proposed method MVCMF and MVCMF-FHALs.

Signature: ***Date:*** 24-06-2019

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all Co-authors confirming their certifying authorship.

Name: Dr Richi Nayak, Associate Professor

Signature: ***Date:*** 24-06-2019

ABSTRACT. The challenge of clustering multi-view data is to learn all latent features embedded in multiple views accurately and efficiently. Existing Non-negative matrix factorization-based multi-view methods learn the latent features embedded in each view independently before building the consensus matrix. Hence, they become computationally expensive and suffer from poor accuracy. We propose to formulate and solve the multi-view data representation by using Coupled Matrix Factorization(CMF) where the latent structure of data will be learned directly from multiple views. The similarity information of data samples, computed from all views, is included into the CMF process leading to a unified framework that is able to exploit all available information and return an accurate and meaningful clustering solution. We present a variable selection based Greedy Coordinate Descent algorithm to solve the formulated CMF to improve the computational efficiency. Experiments with several datasets and several state-of-the-art benchmarks show the effectiveness of the proposed model.

5.1 Introduction

Multi-view clustering has attracted significant attention in recent years due to its effectiveness and applicability to many fields such as web mining and computer vision, where multi-faceted data is naturally generated [35, 103, 120]. An example of multi-view data is a web search system where web-pages can be represented with content or links or search queries or viewers. Another example is a multilingual website where content can be represented in multiple different languages. Multi-view data, where the data can be represented with multiple views, provides complementary and compatible information to a clustering algorithm and assists it to learn accurate and meaningful outcome [116]. Early approaches to cluster multi-view data concatenate the data from all views to obtain a single view and apply traditional clustering methods. This approach includes mutual information from

multiple views in the process, however, it ignores the specific characteristics and the structural information inherent in individual views [46, 118]. In recent years, several methods have been proposed that exploit multi-view representation explicitly and showed improved clustering quality [118]. Amongst them, the most popular and effective methods are matrix factorization-based that learn the embedded space by projecting high-dimensional data to a lower-dimensional space and seek the cluster structures in this new space.

The multi-view clustering methods based on Non-Negative Matrix Factorization (NMF) [64] framework present each view data as a matrix, factorize each matrix independently and find the consensus factor matrix among all views (as shown in Figure 5.1.a). A multi-view clustering algorithm can only perform effectively when it is able to learn the latent features and typical information from multiple views in a consensus manner. Most of the NMF-based multi-view methods [56, 58, 102, 103] first use a collection of matrices to formulate the multi-view problem then learn the latent features embedded in each view. In the later step called integration or fusion, the consensus latent feature matrix is achieved by linearly combining each factorized view matrix or using regularization to guide the consensus factor matrix learning process. The well-known NMF-based method, MultiNMF [56], incorporates the NMF framework with Probabilistic Latent Semantic Analysis (PLSA) constraint [47] to seek the common latent feature matrix from the low representations of all views. The multi-view representation setting in MultiNMF has been extended in many later works. The Adaptive Multi-view Semi-Supervised Non-negative Matrix Factorization (AMVNMF) method [58] uses label information as prior knowledge and uses it as a hard constraint to enhance the discriminating power in MultiNMF. MMNMF [120] incorporates a multi-manifold regularization into the multi-view NMF framework which can preserve the manifold learned from multiple views. Multi-component nonnegative matrix factorization (MC-NMF) [102] and Diver NMF (DiNMF) [103] aim to exploit the diverse information from different data views. This approach

shows effectiveness as it understands the underlying structures in complex datasets by considering all compatible and complementary information from all views for the learning process.

Since the multi-view data contains the data where samples are represented via different features, we conjecture that all the views data should be decomposed jointly and simultaneously to discover the shared features directly from the views, instead of separately decomposing each view and then looking for the common latent representations in the fusion step as in most of the NMF-based multi-view clustering methods. We propose a novel solution by using a Coupled Matrix Factorization (CMF) model [1, 2] to effectively formulate and solve the multi-view representation problem based on the NMF framework [64]. The CMF representation allows us to make a shared mode amongst all views. The goal is to find a distinct base matrix corresponding to each view and the consensus coefficient matrix generated from the shared mode which captures underlying structure of data samples as well as commonality among multiple views.

The NMF-based clustering methods are known to produce a quality solution at the cost of computation time. Majority of NMF methods use Multiplicative Update Rule (MUR) [28] for optimization. MUR has the advantage of good compromise between speed and the ease of implementation [64], however, it has been criticized for failing to guarantee the convergence of component matrices to a stationary point [30]. Recently, many Coordinate Descent (CD) based updating methods, that focus on updating one element at a time until convergence, have been proposed such as HALs or FHALs [25] where the algorithm will sequentially update each element of every updating matrix (HALs) or cyclic update all elements in each matrix before starting to update the others in the same manner (FHALs). Greedy Coordinate Descent (GCD) [49] is proposed to carefully and greedily select elements relying on their importance to update at each step rather than treating every element equally as in HALs or FHALs. GCD outperforms other updating methods in terms of

convergence time as well as it can result in a better lower-order dimensional representation due to its ability to emphasize updating important elements at each iteration. We propose to use the Greedy Coordinate Descent (GCD) method which has been proved effective in traditional matrix completion problem as compared to MUR [49]. In this paper, we propose to extend GCD to solve the multi-view clustering problem.

While projecting the search space from high to lower-dimension, the NMF based methods do not guarantee that the geometric data structure will not change after the lower space mapping. Traditionally, NMF based methods rely on manifold learning to discover and preserve the local structures of data, i.e., to learn and maintain the similarities between data points within local areas [11, 21]. Yet due to multi-view data represented through many views, it is not a trivial task to learn the consensus manifold for multi-view data [66, 120], i.e., learning and preserving the consensus similarity information of all views. In this paper, taking the advantage of CMF, we propose to embed a similarity matrix into the factorization process to be decomposed as a new sample-feature matrix. The similarity matrix is the linear combination of similarity information derived from all views. This novel solution by factorizing the similarity matrix while higher to lower dimension mapping will ensure the consensus matrix to embed all similarity information during optimizing.

The proposed Multi-view Coupled Matrix Factorization (MVCMF) method has been tested with various diverse datasets and benchmarked with state-of-the-art multi-view clustering methods. The experimental results show significant improvement in terms of both accuracy and time complexity. It ascertains that (1) the proposed MVCMF representation is able to exploit the similarity information of data samples from multiple views more effectively, and (2) the proposed optimization method is able to produce a better representation with important elements selection at each optimizing iteration.

More specifically, the contributions of this paper are: (1) Formulate the multi-

view representation problem effectively with CMF to cluster multi-view data; (2) Embed the similarities between samples on each view in CMF which can improve the quality of clustering while keeping the framework simple but comprehensive; (3) Present a single variable update rule of GCD optimization for the proposed unified framework. To our best of knowledge, this is the first work presenting CMF and combining samples similarity into a unified framework to cluster multi-view data. This framework respects data characteristics from multiple views and at the same time respecting similarities between data points within a cluster. This is also a first work where GCD is applied in NMF optimization in multi-view setting.

5.2 The Proposed Multi-view Coupled Matrix Factorization (MVCMF) Clustering Technique

5.2.1 Problem definition - Traditional NMF-based Multi-view Clustering

Suppose $X = \{X_1, X_2, ..., X_{n_v}\}$ is a multi-view dataset with n_v views in total. Data in view v is represented in data matrix $X_v \in R_+^{n \times m_v}$ where n is the number of samples and m_v is the number of features of the vth view.

The NMF-framework to factorize multiple views of data can be written as [56, 102],

$$J_1 = \min \sum_{v=1}^{n_v} \|X_v - H_v W_v\|_F^2, \text{ s.t. } H_v \geq 0, W_v \geq 0 \tag{5.1}$$

where $H_v \in R_+^{n \times k}$ is the new low-rank representation of data corresponding to the basis $W_v \in R_+^{k \times m_v}$ under the vth view and k denotes the number of new rank.

In the fusion step, the consensus latent feature matrix, denoted as H_*, is calcu-

lated by taking the average [103],

$$H_* = \sum_{v=1}^{n_v} H_v / n_v \tag{5.2}$$

or linearly combining [102]

$$H_* = [H_1 ... H_{n_v}] \tag{5.3}$$

The objective function in Eq. (5.1) is able to learn different data representations from different data views. In the later step, the consensus data matrix that embeds the cluster structure of data samples will be learned based on the new learned data representations from all views. Different H_v may contribute different impact on computing H_*. However, since H_* will be computed by taking the average or linear combination of all H_v as in Eqs. (5.2) or (5.3), this will balance the importance of different data views in the latent feature matrix and will fail to ensure the correctness and meaningfulness of the final clustering solution. Moreover, data in multi-view has been well-known of complementary and compatible characteristics, independently learning embedded structures from different views before computing the consensus data will ignore the associative nature among different views during the learning process. Therefore it may not learn good latent features of the multi-view data.

Alternatively, some methods learn the coefficient matrix of each view and simultaneously seek the consensus coefficient matrix by (5.1) minimizing the disagreement between the consensus latent matrix and the new learned latent feature matrices of all views as in Eq.(5.4) [56], or (5.2) looking for the consensus manifold of data as in Eq. (5.5) [120].

$$J_2 = \min \sum_{v=1}^{n_v} \|X_v - H_v W_v\|_F^2 + \sum_{v=1}^{n_v} \|H_v - H_*\|_F^2, \text{ s.t. } H_v \geq 0, H_* \geq 0, W_v \geq 0 \tag{5.4}$$

$$J_3 = \min \sum_{v=1}^{n_v} \left(\|X_v - H_v W_v\|_F^2 + \|H_v - H_*\|_F^2 + \|L_v - L_*\|_F^2 \right) + Tr(H_*^T L_* H_*),$$
$$\text{s.t. } H_v \geq 0,\ H_* \geq 0,\ W_v \geq 0 \tag{5.5}$$

where the intrinsic manifold L_* is the linear approximation of all views manifolds, $L_* = \sum_{v=1}^{n_v} \lambda_v L_v$ and $L_v = D_v - S_v$ is the Laplacian matrix on each view, D_v is the diagonal matrix, $(D_v)_{ii} = \sum_j (S_v)_{ij}$ and S_v is the weighted adjacency matrix [21, 48].

Additional terms in Eqs. (5.4), (5.5) help finding the optimal coefficient matrix by constraining the optimal matrix to be as close to different views as possible or by constraining the final space that lies on a convex hull of the data manifold, thus can bring the more realistic common latent matrix H_* as compared to learning H_* as in Eqs. (5.2) or (5.3). However, the step of constraining H_v to find the optimal latent feature matrix H_* dramatically increases computation as well as it introduces the computation approximation errors that affect the effectiveness of the model. For instance, objective function in Eq. (5.4) results in $\epsilon = \epsilon_1 + ... + \epsilon_{n_v}$, the sum of errors ϵ_v of computing H_* from each H_v.

To overcome these inherent problems, we propose to reformulate the NMF-based multi-view representation by using the CMF model as well as embed the intra-similarities among data objects in each view in the factorization process.

5.2.2 Proposed Multi-view Coupled Matrix Factorization (MVCMF)

We propose to formulate the multi-view representation using CMF as follows,

$$J = \min \sum_{v=1}^{n_v} \|X_v - H_* W_v\|_F^2, \text{ s.t. } H_* \geq 0, W_v \geq 0 \tag{5.6}$$

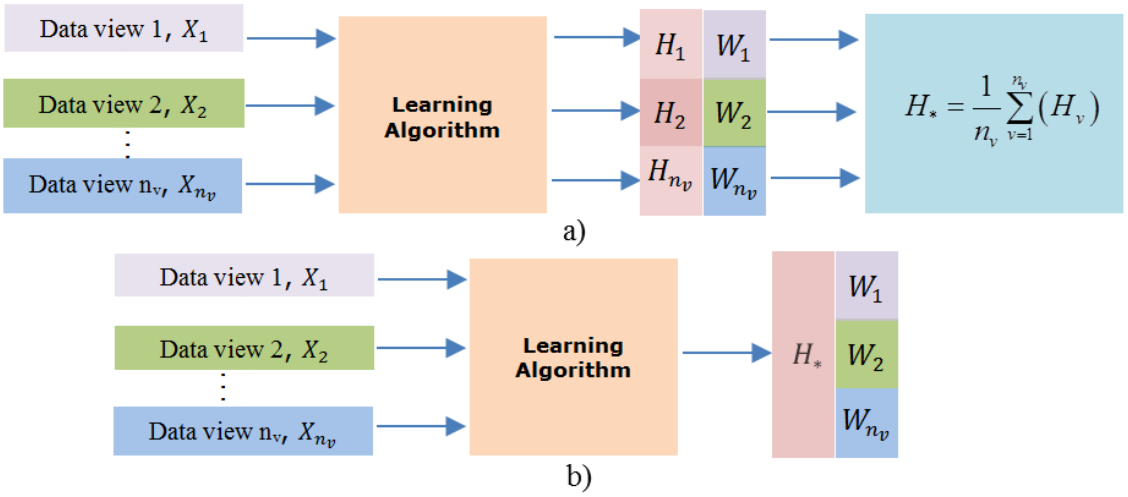


Figure 5.1: Traditional NMF-based versus CMF for Multi-view Learning. H_v encodes the latent features learned from data view vth. W_v is the corresponding basis matrix and H_* is the common consensus matrix.

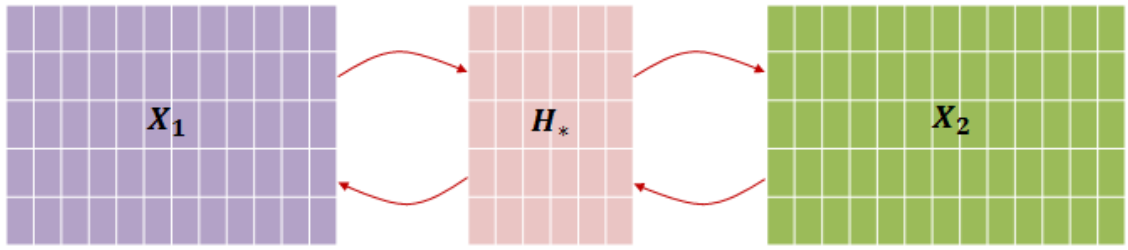


Figure 5.2: The consensus latent matrix H_* is learned by a CMF model in a two-view dataset. H_* is first learned from data view 1, i.e., X_1 and then passed onto data view 2, i.e., X_2 to be updated. H_* is iteratively updated by simultaneously using data matrices from all views until converged.

where $W_v \in R_+^{k \times m_v}$ is the basis representation of data under the vth view and H_* is the consensus coefficient matrix learned from all views. In this objective function, factor matrices to be updated include H_* and $\{W_v\}_{v=1..n_v}$. Since data represented by multiple views share the same mode, the optimal coefficient matrix H_* should be shared among the low-dimensional spaces learned from data of different views. Figure 5.1.b shows how CMF works as compared to traditional NMF-based framework (5.1.a). The latent low-dimensional space of data can be learned by integrating all information from multiple views and by considering the association relationship between different data views. Figure 5.2 shows an example of learning H_* by CMF model that can capture the truly latent features of multi-view data.

By using CMF, we bring the following benefits. (1) Firstly, the number of components to be updated has been decreased by half in CMF as compared to traditional multi-view NMF framework. There are $n_v + 1$ components to be updated in the CMF model, i.e., $\{W_v\}_{v=1..n_v}$ and H_*, as compared to $2n_v + 1$ components in the multi-view NMF model, i.e., $\{W_v\}_{v=1..n_v}$, $\{H_v\}_{v=1..n_v}$ and H_*. This significantly reduces computational time. (2) Secondly, the objective function in Eq. (5.6) ties factors of different views together to capture data as close as to the fine-grained clusters. This results in a unique and stable solution as possible. This is unlike traditional multi-view NMF formulation that is known to produce multiple non-unique solutions because each view is learned independently before seeking the consensus matrix that leads to different solutions of H_*. (3) Lastly, the latent consensus feature matrix is learned naturally since it has been incorporated during the learning process. This reduces approximation errors in contrast to the traditional NMF where they are accumulated as in eqs. (5.4) or (5.5).

We further improve the objective function in Eq. (5.6) by adding the intra-similarities of data samples represented on all views. Consider view v, the similarity matrix S_v of data samples on this data view X_v is constructed using k nearest-neighbour graph (kNN) as in [21, 48, 72]. Specifically,

$$S_v(x_i, x_j) = \begin{cases} s_v(x_i, x_j) \text{ if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{5.7}$$

where $\mathcal{N}_k(x_i)$ denotes the kNN points of x_i and $s_v(i, j)$ is the weight of the affinity between x_i and x_j on view v. The affinity value can be represented as Binary, Heat Kernel, Cosine or Dot-Product [21]. We have different similarity matrices S_v of size $n \times n$ computed from different views showing how data samples are similar to each other on each view. We use a linear combination to combine the affinity information

from these matrices into a single matrix as follows,

$$S = \sum_{v=1}^{n_v} \lambda_v S_v \tag{5.8}$$

We conjecture that S is able to encode different view semantics by including how close each sample is to others to an extent. λ_v is the parameter that can be tuned to set different levels of contributions of different views to the learning process. In this work, we aim to treat all views equally hence the parameters will be set to $1/n_v$ in our experiment. Regarding S as a new data view and embedding in the objective function J will ensure the learning process to find clusters where closeness between objects is high within each cluster.

The objective function J from Eq. (5.6) can be written as follows,

$$J = \min \sum_{v=1}^{n_v} \|X_v - H_* W_v\|_F^2 + \|S - H_* B\|_F^2, \text{ s.t. } H_* \geq 0,\ B \geq 0,\ W_v \geq 0 \tag{5.9}$$

B in Eq. (5.9) is the corresponding basis matrix when factorizing the similarity matrix S.

We also note that this objective function can be modified to add various constraints such as $l1$ norm, $l2$ norm or orthogonal constraint depending on the property of dataset or the objective of learning algorithm, i.e., achieving sparseness or even distributions or uniqueness for the factor matrices. There exists several NMF methods considering these constraints [48, 29]. We omit them and aim at introducing a fundamental and novel coupled-matrix framework for clustering multi-view data where all available information has been exploited.

5.2.3 The Proposed Optimization solution

Existing NMF-based methods apply MUR [64] to solve optimization where every element in the matrix is equally updated in each iteration. Distinct from these

methods, we conjecture that more important elements should be updated more frequently to learn a better low-order representation.

We propose to solve the objective function in Eq. (5.9) using GCD so that the important elements can be selected and important features for each factor matrix can be learned. The factor matrices to be updated from Eq. (5.9) are H_*, B and $\{W_v\}_{v=1..n_v}$. The proposed updating framework will switch between the factor matrices similar to other CD based methods [49, 26]. Specifically, we have the following outer updating,

$$(W_1^0, W_2^0, ..., W_v^0, B^0, H_*^0) \rightarrow (W_1^1, W_2^0, ..., W_v^0, B^0, H_*^0)$$

$$\rightarrow ... \rightarrow (W_1^n, W_2^n, ..., W_v^n, B^n, H_*^n)$$

and the following inner updating,

$$(W_1^i, W_2^i, ..., W_v^i, B^i, H_*^i) \rightarrow (W_1^{i,2}, W_2^i, ..., W_v^i, B^i, H_*^i)$$

GCD is a row-based updating scheme where each inner iteration will iteratively update variables in the ith row until the inner stopping condition is met. In order to solve objective function J utilizing GCD, we need to solve H_*, B and $\{W_v\}_{v=1..n_v}$ via the below updating solutions.

Proposed Rigorous Single Variable Update Rule

We propose the rigorous one-variable sub problem for the GCD update rule to solve the MVCMF objective function in Eq. (5.9). We explain the update rule for H_* while it can be derived for other factor matrices similarly. To update a single variable in the shared factor matrix H_*, the primary goal is to fix all the other factor matrices and modify the shared factor matrix by adding a scalar value s to the selected single variable. This scalar value implies how much the single variable

value should be tuned to achieve minimal objective difference. It can be expressed as,

$$(H_*, W_v, B) = (H_* + sZ, W_v, B) \tag{5.10}$$

where Z is a zero matrix except the element, that is being updated, is set to 1.

In order to identify the value of s, the rigorous one-variable sub problem is defined,

$$\min_{s:H_*+s>0} g^{H_*}(s) \equiv f(H_* + sZ, W_v, B) \tag{5.11}$$

Above equation can be rewritten as,

$$g^{H_*}(s) = g^{H_*}(0) + g'^{H_*}(0)s + \frac{1}{2}g''^{H_*}(0)s^2 \tag{5.12}$$

Where g' and g'' are the first and second order derivatives of H_* representing the gradient and Hadamard product respectively. For simplicity we have

$$g' = \partial J/\partial H_* = G^{H_*} \tag{5.13}$$

$$g'' = \partial^2 J/\partial H_* = H^{H_*} \tag{5.14}$$

The update rule to identify the value of s that is added to H_* is

$$s = \frac{g'}{g''} = \frac{G^{H_*}}{H^{H_*}} \tag{5.15}$$

Using the one variable sub problem and respective gradients, the variable importance (Magnitude Matrix D^{H_*}) as the difference in objective function can be calculated as follows,

$$D^{H_*} = g^{H_*}(0) - g^{H_*}(s) \tag{5.16}$$

By substituting Eq. (5.12) in Eq. (5.16), the variable importance equation is as,

$$D^{H_*} = -(H_* * G^{H_*}) - \frac{1}{2} * (H^{H_*} * s^2) \tag{5.17}$$

Updating solution for H_*

Taking the first and the second derivative of J corresponding to H_* we have:

$$\partial J/\partial H_* = -2\sum_v X_v W_v^T + 2H_* \sum_v W_v W_v^T - 2SB^T + 2H_* BB^T \tag{5.18}$$

$$\partial J^2/\partial H_* = 2\sum_v W_v W_v^T + 2BB^T \tag{5.19}$$

According to the CD method [25], H_* will be updated through the update rule,

$$(H_*)_{ir} = [(H_*)_{ir} + (\hat{H}_*)_{ir}]_+ \tag{5.20}$$

where the $[\ ]_+$ notation is for non-negative constraint and the additional value $(\hat{H}_*)_{ir}$ to be added in element $(H_*)_{ir}$ at each update step is computed as,

$$(\hat{H}_*)_{ir} = (H_*)_{ir} - \frac{\partial J/\partial H_*}{\partial^2 J/\partial H_*} \tag{5.21}$$

setting,

$$G^{H_*} = \partial J/\partial H_* = -2\sum_v X_v W_v^T + 2H_* \sum_v W_v W_v^T - 2SB^T + 2H_* BB^T \tag{5.22}$$

For simplicity, let us represent

$$Ha^{H_*} = \frac{\partial^2 J}{\partial H_*} \tag{5.23}$$

With the pre-calculated gradient and Hadamard product, the magnitude matrix

$D^{H_*} = \{D^{H_*}\}_{ir}^{n\times k}$, is calculated where $(D^{H_*})_{ir}$ encodes how much objective function value will be decreased when each corresponding element $(H_*)_{ir}$ is updated.

$$(D^{H_*})_{ir} = -G_{ir}^{H_*}(H_*)_{ir} - 0.5Ha^{H_*}(H_*)_{ir}^2 \tag{5.24}$$

Based on this magnitude matrix, for each i^{th} row, the most important element is updated. By updating the gradient of all the elements in i^{th} row, we can identify the next most important element to be updated by recalculating the i^{th} row of magnitude matrix. This process is repeated until the inner stopping condition or maximum inner iteration is reached. The working process is explained in Algorithm 3.

Updating solution for W_v

Taking the first and the second derivative of J in Eq. (5.9) corresponding to W_v we have,

$$\partial J/\partial W_v = -2X_vH_*^T + 2H_*H_*^TW_v \tag{5.25}$$

$$\partial J^2/\partial W_v = 2H_*H_*^T \tag{5.26}$$

W_v will be updated through the update rule

$$(W_v)_{ir} = [(W_v)_{ir} + (\hat{W}_v)_{ir}]_+ \tag{5.27}$$

where the additional value $(\hat{W}_v)_{ir}$ to be added into element $(W_v)_{ir}$ at each update step is computed as,

$$(\hat{W}_v)_{ir} = (W_v)_{ir} - \frac{\partial J/\partial W_v}{\partial^2 J/\partial W_v} \tag{5.28}$$

We define the magnitude matrix $D^{W_v} = \{D^{W_v}\}_{ir}^{m_v\times k}$, where $(D^{W_v})_{ir}$ encodes how much objective function value will be decreased when each corresponding element

Algorithm 3: The Proposed MVCMF Algorithm

Input : Multi-view data matrices $\{X_v\}$; Randomly initiated factor matrices; $D = \emptyset$; Rank R; $init, inner\_iter, tol, avalue$.

Output: Factor matrices: H_*, $\{W_v\}_{v=1..n_v}$ and B.

Compute G^{H_*}, Ha^{H_*} using Eqs. (5.22), (5.23).
for *each* i, r, $1 \leq i \leq I$, $1 \leq r \leq R$ **do**
  $(\hat{H}_*)_{ir} \leftarrow G^{H_*}_{ir} / Ha^{H_*}_{ir}$
  $(\hat{H}_*)_{ir} \leftarrow (H_*)_{ir} - (\hat{H}_*)_{ir}$
  $(\hat{H}_*)_{ir} = \max((\hat{H}_*)_{ir}, 0)$
  Update $(\hat{H}_*)_{ir}$ and $(D^{H_*})_{ir}$ as in Eqs. (5.21), (5.24)
  $init = \max((D^{H_*})_{ir}, init)$
end
for *each* p, $1 \leq p \leq I$ **do**
  while *inner stopping condition* **do**
    $q = -1, bestvalue = 0$
    for *each* r, $1 \leq r \leq R$ **do**
      $(\hat{H}_*)_{pr} \leftarrow G^{H_*}_{pr} / Ha^{H_*}_{pr}$
      $(\hat{H}_*)_{pr} \leftarrow (H_*)_{pr} - (\hat{H}_*)_{pr}$
      $(\hat{H}_*)_{pr} = \max((\hat{H}_*)_{pr}, 0)$
      Update $(\hat{H}_*)_{pr}$ and $(D^{H_*})_{pr}$ as in Eqs. (5.21), (5.24)
      if $(D^{H_*})_{pr} > bestvalue$ **then**
        $bestvalue = (D^{H_*})_{pr}$, $q = r$
      end
    end
    break if $q = -1$
    $(H_*)_{pr} \leftarrow (H_*)_{pr} + (\hat{H}_*)_{pr}$
    $G^{H_*}_{pr} = G^{H_*}_{pr} + ((\hat{H}_*)_{pr} * Ha^{H_*}_{qr})$ **for all** $r = 1..R$
    break if $(bestvalue \leq init * tol)$
  end
end
For updates to W_v and B, repeats analogues these above steps.

$(W_v)_{ir}$ is updated as below,

$$(D^{W_v})_{ir} = -G^{W_v}_{ir}(W_v)_{ir} - 0.5\frac{\partial^2 J}{\partial W_v}(W_v)^2_{ir} \tag{5.29}$$

Updating solution for B

Similar to updating schemes for H_* and W_v, B will be updated through the update rule,

$$(B)_{ir} = [(B)_{ir} + (\hat{B})_{ir}]_+ \tag{5.30}$$

The additional value $(\hat{B})_{ir}$ to be added into element $(B)_{ir}$ at each update step is computed as,

$$(\hat{B})_{ir} = (B)_{ir} - \frac{\partial J/\partial B}{\partial^2 J/\partial B} \tag{5.31}$$

The magnitude matrix $D^B = \{D^B\}_{ir}^{n\times k}$ is calculated as below,

$$(D^B)_{ir} = -G^B_{ir}(B)_{ir} - 0.5\frac{\partial^2 J}{\partial B}(B)^2_{ir} \tag{5.32}$$

where $G^B = \partial J/\partial B = -2SH_*^T + 2H_*H_*^T B$.

It can be noted that the steps of iteratively updating all important elements on each row is equivalent to selecting important features to be updated instead of equally updating all elements as in MUR or FHALs [25, 28]. By these single variable update rules, the factor matrices can be updated as per Figure 5.1.b that avoids the frequent update of the shared matrix H_* and reduces the learning complexity.

5.3 Experiments and Results

The proposed MVCMF method is evaluated using several multi-view datasets exhibiting different sizes and different numbers of views (as detailed in Table 5.1). Movie (D1) is extracted from the IMDB Website[1] that contains a set of movies represented by two different views that are actors and keywords. Cora[2] (D2) is a scientific publication dataset with documents represented via two views, content and cites. Two multi-view benchmark datasets R-MinMax (D3) and R-Top (D4) are selected from the Reuters-21578 [3], a well-known text dataset. R-MinMax (D3) includes 25 classes of the dataset with at least 20 and at most 200 document per class and R-Top (D4) contains 10 largest classes in the dataset. Apart from the first view where documents are represented via terms, we used external knowledge, i.e.,

[1] http://www.imdb.org
[2] http://www.cs.umd.edu/%7Esen/lbc-proj/data/cora.tgz
[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Wikipedia, to represent the second view of components (following steps as in [48]). Two subsets MLR-36k4k (D5) and MLR-84k5k (D6) were extracted from the most well-known multi-view dataset of Reuters RCV1/ RCV2 Multilingual with English, French, Italian and Germany terms representing each view.

We utilize two popular measurement criteria, accuracy which measures the percentage of correctly obtained labels and Normalized Mutual Information (NMI) [119] which measures the purity against the number of clusters. We also investigate the runtime performance. Average results are reported after 5-fold runs of each experiment.

We empirically evaluate the performance of MVCMF against the following state-of-art methods.

1. ConcatK-Means: In this method, we concatenate the features of all views to create a single view data and apply K-means to achieve the clustering result.

2. ConcatNMF: In this method, we concatenate the features of all views to create a single view data and apply the traditional NMF [28].

3. ConcatGNMF: Data features from all views of dataset are concatenated before executing GNMF [21].

4. MultiNMF [56]: The method simultaneously learns the low-dimensional representations from all views before seeking the consensus low-dimensional matrix.

5. MVCMF-FHALs: This is variation of our proposed method MVCMF with the FHALs updating scheme [25], to check the effectiveness of using GCD in clustering problem.

5.3.1 Performance analysis

As illustrated in Tables 5.2-5.3, MVCMF outperforms other state-of-the-art benchmarking methods on all datasets. This performance is exhibited due to the fact that the

Table 5.1: Characteristic of the datasets

Properties	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
$\neq$ Classes	17	7	25	10	6	6
$\neq$ views	2	2	2	2	4	4
$\neq$ Samples	617	2,708	1,413	4,000	3,600	8,400
$\neq$ features of view 1	1,398	1,433	2,921	6,000	4,000	5,000
$\neq$ features of view 2	1,878	5,429	2,437	6,000	4,000	5,000
$\neq$ features of view 3	-	-	-	-	4,000	5,000
$\neq$ features of view 4	-	-	-	-	4,000	5,000

Table 5.2: Accuracy of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
ConcatK-Means	23.18	54.91	53.86	52.20	37.61	49.11	45.15
ConcatNMF	28.20	53.77	56.33	37.78	44.81	49.18	45.01
ConcatGNMF	23.99	23.23	42.18	52.70	23.86	31.49	32.91
MultiNMF	23.01	46.16	52.80	47.35	42.00	39.95	41.88
MVCMF-FHALs	27.23	39.07	56.55	53.20	45.75	**53.07**	45.81
MVCMF-GCD	**33.71**	**58.68**	**67.02**	**57.67**	**48.42**	49.61	**52.52**

Table 5.3: NMI of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
ConcatK-Means	20.86	35.59	68.92	**50.35**	21.23	31.92	38.15
ConcatNMF	28.56	34.48	64.51	37.96	30.18	**32.15**	37.97
ConcatGNMF	24.22	05.11	55.07	41.49	09.88	08.22	24.00
MultiNMF	20.87	30.49	66.24	43.07	23.39	21.49	34.26
MVCMF-FHALs	28.46	15.83	69.40	40.38	25.76	30.28	35.02
MVCMF-GCD	**33.96**	**37.96**	**72.22**	46.77	**34.89**	31.68	**42.92**

Table 5.4: Running time (in seconds) of each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**Average**
ConcatK-Means	**0.49**	12.16	6.28	68.77	51.77	187.89	54.54
ConcatNMF	1.73	7.06	6.56	29.38	36.26	103.59	30.76
ConcatGNMF	1.77	7.12	6.63	29.69	34.43	103.92	30.59
MultiNMF	31.01	157.47	120.50	724.29	790.81	2377.00	701.18
MVCMF-FHALs	1.03	**0.86**	**0.86**	**7.57**	**0.58**	13.25	**4.02**
MVCMF-GCD	8.03	1.39	1.06	7.91	3.34	**8.70**	**5.98**

optimal latent features matrix of multi-view data has been learnt jointly, simultaneously and directly from all views data. In addition, the similarity matrix embedded in the factorization helps MVCMF to exploit the similarity information of data samples from multiple views, therefore, resulting in the meaningful solution. More importantly, on datasets D1-3 and D5, MVCMF obtains significantly better clustering results in terms of both accuracy and NMI as compared to other benchmark methods. Though ConcatK-Means and ConcatNMF outperform MVCMF-GCD on D4 and D6 in terms of NMI, the difference is negligible. This validates the effectiveness of our approach of using coupled matrix for multi-view representation utilising GCD framework.

While exploiting the same amount of information, i.e., data from all views, MultiNMF achieves better performance as compared to ConcatNMF and ConcatGNMF. It is because MultiNMF simultaneously learns factor matrices from all views and seeks the consensus latent features instead of concatenating all features and applying single view method as the two methods ConcatNMF and ConcatGNMF do. Similarly, K-means fails to bring a good result on multi-view data, though it is the most well- known method for clustering one view data.

It can also be observed that ConcatGNMF gives a very low performance on high dimensional datasets D2, D5, D6. This method relies on manifold learning to discover the latent structures for data. Unfortunately, it is a non-trivial task of learning the embedded manifold of data based on the kNN graph. Since the kNN will become meaningless as the difference between distances from a point to its nearest and farthest points becomes non-existent in high-dimensional data [13], a poor performance is produced.

Comparing MVCMF-FHALs and MVCMF-GCD with only difference of the proposed optimization method, it can be ascertained by the improved clustering results on all datasets that the proposed optimization method is able to learn a better lower-order representation.

With regard to time complexity, it can be seen from Table 5.4, the proposed MVCMF easily outperforms other benchmark methods. MVCMF-GCD is 11 to 272 times faster than the existing methods for bigger datasets. The reason behind the better runtime performance is that the coupled matrix factorization avoids updating the same factor matrices multiple times and equally updates all the factor matrices. We also note that for small dataset D1, concatenation based methods are faster, but the runtime increases exponentially with increase in the data size, unlike MVCMF. Due to the frequent gradient updates that helps to learn the factor matrices more accurately, the runtime of GCD is higher than FHALs. In some cases like D4 and D6, it can be noted that GCD outperforms or performs equally with FHALs due to the fast convergence achieved through the single variable selection strategy.

5.4 Conclusion

In this paper, we have presented the novel unified coupled matrix factorization framework that incorporates similarity information of data samples calculated from all views to efficiently generate clusters. We also present the single variable update rule of GCD for the proposed framework which exploits variable importance strategy to accurately learn the factor matrices that improves clusters quality. Experiment results on many datasets with different sizes and different number of views show the effectiveness of the proposed framework MVCMF and learning algorithm MVCMF-GCD. In future work, we will further investigate to extend the proposed framework in tensor factorization to handle higher order multi-view clustering.

Paper 6. Khanh Luong and Richi Nayak, *Learning the Consensus and Complementary Information for Multi-View Data Clustering.* In 21st International Conference on Computer Vision - ICCV 2019 (Under Review).

Statement of Contribution of Co-Authors

The authors of the papers have certified that:

1. They meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;

2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;

3. There are no other authors of the publication according to these criteria;

4. Potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and

5. They agree to the use of the publication in the student's thesis and its publication on the QUT ePrints database consistent with any limitations set by publisher requirements.

Contributors:

Khanh Luong (PhD Candidate): Conceived the idea, designed and conducted experiments, analysed data, wrote the paper and addressed reviewers comments (for published papers) to improve the quality of paper.

Signature: ***Date:*** 24-06-2019

A/Prof. Richi Nayak: Provided critical comments in a supervisory capacity on the design and formulation of the concepts, method and experiments, edited and reviewed the paper.

Signature: ***Date:*** 24-06-2019

ABSTRACT. We design a factorization-based loss function to simultaneously learn two components encoding the consensus and complementary information for multi-view data, respectively, by using both Coupled Matrix Factorization (CMF) and Non-negative Matrix Factorization (NMF). We propose a novel optimal manifold for multi-view data which is the most consensed manifold embedded in the high-dimensional multi-view data. This newly optimal manifold is learned and incorporated in the proposed factorizing-based framework. A new complementary enhancing term is added in the loss function to include all possible consensus and complementary information inherent in the multi-view data. An extensive experiment with diverse datasets, benchmarking the state-of-the-art multi-view clustering methods, has demonstrated the effectiveness of 2CMV.

5.5 Introduction

Multi-view data, the data represented by multiple types of features or collected from multiple sources as shown in Figure 5.3, has become ubiquitous and insightful. Multi-view data learning is an emerging topic of interest due to its capacity to embed the rich information distinctly in the learning process, instead of just concatenating the multiple types of features in a single matrix. Multi-view data learning methods bring benefits to many fields ranging from text mining to bioinformatics to computer vision [92, 94].

Multi-view data is deemed to contain consensus and complementary information (as shown in Figure 5.4) for learning a meaningful and informative output as compared to the traditional one-view data [73, 107]. The consensus information is expressed in the manner of how different data views embed a compatible (or correlated) latent structure within the data. The complementary information is expressed in the manner of how each view provides a diverse (or distinct) characteristic within the data. Several multi-view data learning methods have been proposed and vary

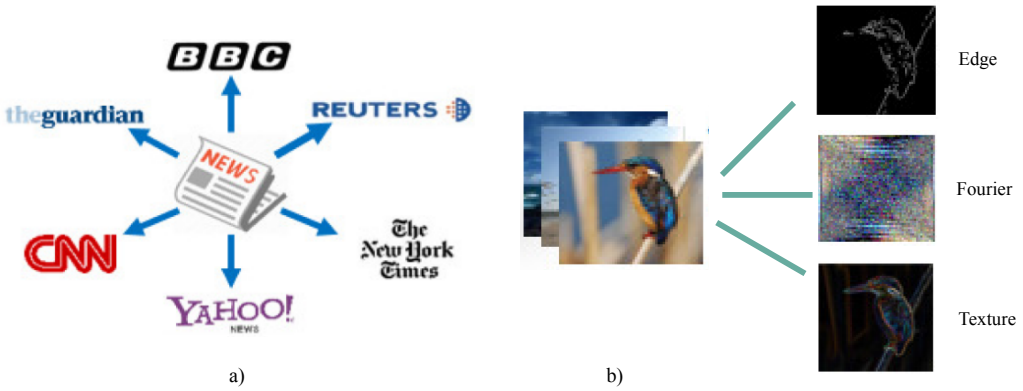


Figure 5.3: Examples of Multi-view Data

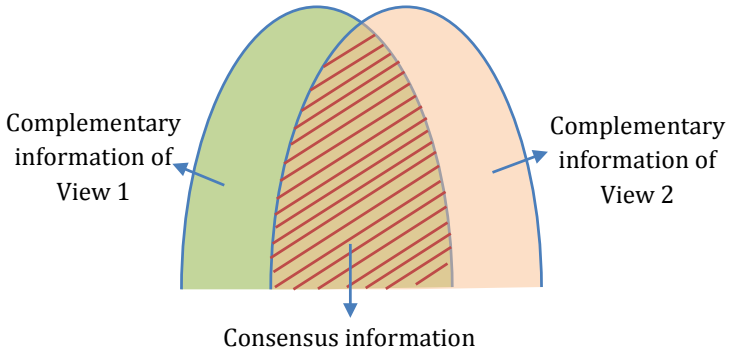


Figure 5.4: Consensus and Complementary information

according to how they use these two characteristics of multi-view data [73, 107]. Since the data characteristics are unknown, it is a non-trivial task to sufficiently learn both consensus and complementary information from multi-view data.

Existing methods focus on learning and/or enhancing either the consensus or the complementary information. For example, MultiNMF [56] learns the consensus representation by embedding the most common latent features learned from all data views. MVSC-CEV [90] learns the common largest eigenvectors from multiple views to produce a cluster solution. MVCMF [71] learns the common representation shared in multiple views directly by using the coupled matrix factorization. These methods focus on learning the most correlated latent features from all views, but they disregard the diverse information present in multiple views and lose the valuable complementary information.

On the other hand, a few methods [23, 103] have been developed that impose a

diversity constraint to explore the complimentary information present in multi views; however, they fail to ensure that the consensus information presents in the data can also be sufficiently learned. Learning consensus and complementary information effectively aids the clustering tasks and produces a meaningful outcome. To the best of our knowledge, no method exists that can learn both the consensus and complementary information adequately.

The overarching aim of this paper is to efficiently and effectively learn both consensus and complementary information to produce an accurate clustering solution for multi-view data. This is done through the following contributions.

First, we design a novel matrix factorization (MF)-based objective function by simultaneously and effectively using the Non-negative Matrix Factorization (NMF) and Coupled Matrix Factorization (CMF) frameworks. NMF [64] has been shown as an effective framework to learn the low-rank representation for clustering since most multi-view data is of high-dimensional and sparse nature [21]. NMF becomes a natural choice to learn the complementary information. Recently, the CMF [1] has been shown as an effective framework for multi-view data learning by allowing a shared mode amongst multiple views [71]. The shared factor matrix is believed to naturally embed the compatible information of multi views. We suppose that the newly learned low-rank representation includes two components, one to embed the good consensus information and another to encode the good complementary information. The proposed objective function is able to naturally learn both types of information simultaneously. *This is our first contribution.*

Second, while projecting the data to lower order in MF, it is necessary to maintain the geometric structure of the data for meaningful representation of the data. Prior research [21, 120] has shown the importance of incorporating manifold learning [11] in the NMF process. The integrated framework ensures the low-rank representation to be smooth with including the intrinsic geometric shape of the original data and enhances the clustering performance [21]. Most existing methods [85, 115],

when combining manifold learning with the factorization based framework designed for multi-view data, rely on building and maintaining the manifold of each view separately. They ignore the correlation between manifolds; however, the intrinsic manifold of multi-view data is believed to sample on a consensus manifold of all views' manifolds. To the best of our knowledge, MMNMF [120] is the only existing method that learns the consensus manifold as a linear combination of all manifolds. A linear manifold may not be sufficient to maintain the geometric shape of the data. Each view manifold may have a different level of contribution to the consensus manifold. Besides, it relies on the consensus coefficient matrix to respect the consensus manifold that is highly error-prone.

In this paper, we introduce a novel concept of optimal manifold for the multi-view data. We propose to build and maintain the optimal consensus manifold to encode all the common closeness information present in multiple views. We integrate the optimal manifold into the two-component factorization-based framework designed in the first contribution, which will ensure that the accurate representation will be learned while the complementary information will be fully exploited and the consensus information will be successfully enhanced. *This is our second contribution.*

Third, as stated before, only a handful of multi-view methods exist that focus on learning the complementary information. The existing NMF-based method, DiverNMF [103] applies a simple diversity constraint on the low-rank representations of different views. However, as shown in experiments later, it fails to capture all diverse information on multi views. Subspace-based methods [23, 112] propose to use the complex constraint such as *Hilbert-Schmidt Independence Criterion* [41] to identify distinct features in the subspace with the aid of kernel learning. We propose a new *complementary enhancing* term based on the concept of orthogonality to highlight the distinctness in the data in the MF framework. It is applied on both components (the low-rank representations based on consensus and complementary information) to encourage the independence between them. In this way, the consensus component

will embed the most consistent information and the complementary component will embed the most distinct information. *This is our third contribution.*

This paper presents the proposed method, aiming to fully exploit the **C**onsensus and **C**omplementary information for **M**ulti-**V**iew data (**2CMV**). We conjecture that 2CMV learns the valuable insightful information, which is the premise of an efficient clustering method. An extensive experiment on several real-world multi-view datasets shows the effectiveness of 2CMV in learning an accurate representation. 2CMV is shown to outperform the state-of-the-art clustering methods.

5.6 Proposed Method: 2CMV

5.6.1 Problem Definition

Suppose $X = \{X_1, X_2, ..., X_{n_v}\}$ is a multi-view dataset with n_v views in total. The data in vth view is represented by the data matrix $X_v \in \mathbb{R}_+^{n \times m_v}$ where n is the number of samples and m_v is the number of features of the vth view.

The factorization based multi-view data clustering task is to group data samples into clusters via a process of simultaneously factorizing all view data to learn a consensus and complementary low-rank representation and then seek a cluster structure from the representation. The more meaningful the lower-rank representation is, the more meaningful the cluster structure is. The goal of factorization based multi-view data learning is to obtain an accurate low-rank representation that embeds the latent features of the (original) high-dimensional and sparse data. Next we will present the three subsections corresponding to the three contributions suggested to build the 2CMV method.

5.6.2 Factorization-Based Loss Function to Simultaneously Learn the Consensus and Complementary Information

The NMF framework [63, 64] and recently, CMF [1] framework, have been popularly but separately used in learning the low-rank representation for multi-view data [71, 102, 103, 115, 120]. We design a novel multi-view factorization-based framework to learn the lower-order representation incorporating both the consensus and complimentary information. The proposed factorization-based objective function, named as the two-component model, simultaneously using both NMF [63, 64] and CMF [1], can be expressed as follows,

$$\begin{aligned} J &= \min \sum_{v=1}^{n_v} \|\mathbf{X}_v - [H_* \; H_v] W_v^T\|_F^2, \\ &\text{s.t. } H_* \geq 0, H_v \geq 0, W_v \geq 0, H_{com} = \sum_{v=1}^{n_v} \frac{H_v}{n_v} \\ H_{final} &= [\alpha H_* \; (1-\alpha) H_{com}], 0 < \alpha < 1 \end{aligned} \tag{5.33}$$

where $H_* \in R_+^{n \times r}$, $H_v \in R_+^{n \times r}$, $W_v \in R_+^{m_v \times r}$, $H_{final} \in R_+^{n \times 2r}$ and r denotes the number of the new low-order rank. The role of each factor matrix in Eq. (5.33) can be explained as follows. H_* is the low-rank representation learned from the shared mode of all views utilising the CMF model. Due to being learned directly and naturally from all view data, H_* is supposed to embed the important compatible latent features from multiple views. H_* acts as the **consensus component** in the proposed objective function.

H_v is the new low-rank representation of data corresponding to the basis W_v of the vth view following the NMF model. Due to being learned directly from each data view v, each H_v carries distinct information present in the view. H_{com} is the **complementary component** matrix designed to combine the low-rank representations

of all views.

A general architecture of the newly designed factorization-based model is shown in Figure 5.5.a. Both NMF and CMF will be utilised to factorize the input multi-view data in order to obtain two components. The NMF model learns new low-rank representation H_v corresponding to the basis W_v under the vth view. These learned H_v, $\forall v = 1..n_v$ create the complementary low-rank representation H_{com} for the two-component model. Simultaneously, the CMF model learns the consensus low-rank representation H_* for the two-component model, corresponding to different basic matrices W_v from different views.

The final low-rank representation H_{final} is a weighted concatenation of the two components (as illustrated in Figure 5.5.b). For the unsupervised learning task, we have no information about the amount of consistent or complementary information present in the dataset. We use the trade-off parameter α in Eq. (5.33) with the constraint $0 < \alpha < 1$ in order to ensure both consistent and complementary components simultaneously contributing to H_{final} and obtaining the best clustering result.

During the optimization process, as per the objective function, H_*, different W_v and H_{com} will simultaneously be updated and the optimal final solution H_{final} representing the multi-view data in the lower-order will be returned.

5.6.3 Learning the Optimal Manifold

The manifold assumption supposes that data is lying on an intrinsic manifold that is embedded in high-dimensional data [96]. The manifold of data can be understood as the geometric structure of data generated by data points that residing close together in space [11].

In multi-view data, it is believed that, data on different views will sample on different manifolds [120, 115]. Moreover, due to embedding the compatible and complementary information, a recent research has shown that there exists the consensus manifold in multi-view data [120]. The consensus manifold is embedded in

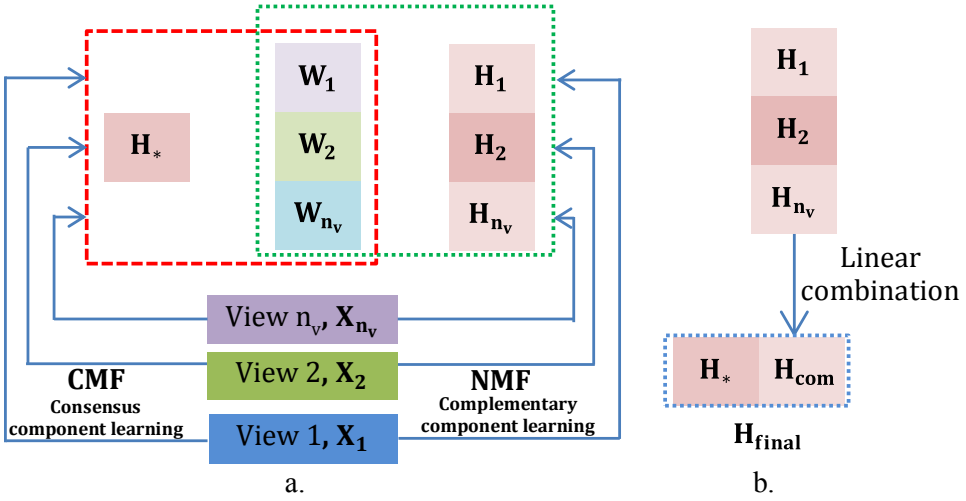


Figure 5.5: The consensus and complementary learning model for multi-view data.

the convex hull, which can be the minimal convex set containing all manifolds or a convex combinations of all manifolds [37]. This leads us to the assumption that, there exists an optimal manifold in multi-view data that can be learned. We define the optimal manifold for multi-view data as the most consensed manifold wherein data points belong to this optimal manifold if and only if they exist in all views' manifolds.

Figure 5.6 illustrates an example of the optimal manifold L_{opt}. Suppose we have a three-view dataset where data from three views generate three different manifolds illustrated as different planes on space. The optimal manifold, in this case, is the intersection part of the three manifolds where the distance property between data points residing on this optimal manifold will be similar on all views' manifolds. The optimal manifold is constructed in the form of a graph regularization [11, 21] as below,

$$L_{opt} = D_{opt} - A_{opt} \tag{5.34}$$

with the optimal affinity matrix from all views, $A_{opt} = \{A_{opt}(i,j)\}$ is constructed as,

$$A_{opt}(i,j) = \min\{A_v(i,j)\}_{v=1..n_v} \tag{5.35}$$

A_v is the affinity matrix corresponding to data view vth, populated using a k nearest neighbour (kNN) graph as in [11],

$$A_v(i,j) = \begin{cases} a_{ij} \text{ if } (x_j \in \mathcal{N}^k(x_i) \text{ or } x_i \in \mathcal{N}^k(x_j) \\ 0, \text{ otherwise} \end{cases} \tag{5.36}$$

where $\mathcal{N}^k(x_i)$ denotes k nearest neighbours of x_i, a_{ij} is the similarity between x_i and x_j. Different similarity measures such as Binary, Heat Kernel, Cosine or Dot-Product can be used [10].

D_{opt} is a diagonal matrix calculated as $D_{opt}(i,i) = \sum_j A_{opt}(i,j)$.

It is obvious that A_{opt} constructed as in Eq. (5.35) is an affinity matrix encoding the most common closeness information of all views, i.e., $A_{opt}(i,j) > 0$ *if and only if* $A_v(i,j) > 0, \forall v = 1..n_v$.

L_{opt} constructed as in Eq.(5.34) is obviously a Laplacian matrix. We are now ready to propose the following graph regularization into the objective function in order to ensure that the sampling of data on the new low dimensional space respects the new optimal manifold,

$$\min Tr(H_v^T L_{opt} H_v), \forall v = 1..n_v \tag{5.37}$$

We have the following lemma,

Lemma 1: The optimization problem in Eq. (5.37) will ensure that all data points residing close in all views in the original space will be preserved to remain close in the new mapped space.

Proof: Consider two data points x_i, x_j with their new low-rank representations h_i, h_j, respectively, on view vth, Eq. (5.37) will be equivalent to the following equation [11],

$$\min \sum_{i,j=1}^{n} \|h_i^v - h_j^v\|^2 A_{opt}(i,j), \forall v = 1..n_v \tag{5.38}$$

If two data points x_i, x_j are close in all views, i.e., $A_v(i,j) > 0$, $\forall v = 1..n_v$, we have $A_{opt}(i,j) > 0$, following Eq. (5.35). When $A_{opt}(i,j) > 0$, minimizing Eq. (5.38) will minimize $\|h_i^v - h_j^v\|^2$, $\forall v = 1..n_v$. It will make the new low-rank representation h_i to close to h_j in the new low dimensional space learned from all views. Note that when $A_{opt}(i,j) = 0$, i.e., x_i and x_j are not residing on the optimal manifold, the minimizing term in Eq. (5.37) will bring no effect to the learning process.

Learning and preserving an optimal manifold will ensure that, (1) the optimal affinity matrix A_{opt} is supposed to encode the common closeness information of data points of all views; (2) two data points will reside close in the optimal manifold if and only if they are close on all views' manifolds; and (3) imposing the optimal manifold on all view data will help in promoting learning the consensus geometric information present in the data.

It can be argued that, since the learned optimal manifold is the most consensed manifold, many data points that do not belong to the manifold will be left uncontrolled and may affect the quality of clusters. In fact, the optimal manifold (learned in this manner) will provide the freedom for the proposed framework to learn the complementary information, since all other data points not residing on the optimal manifold should already be taken care of by the factorization process and the other constraints in our model.

Why it works? There exit two common approaches of using manifold learning in multi-view data [73]. The first approach allows the manifold leaning process to learn the latent features of data represented on each view [48, 66, 72, 115]. The close data points on each view will be forced to be close in the new mapped low-dimensional space. On multi-view data, this process is equivalent to separately learning from distinct views and disregarding the associative relationship among views. This process will not work well if the intrinsic local geometric structures are different from all views. Specifically, if two data points accidentally reside close on one view, they will be forced to be close on its corresponding space while they

actually belong to different clusters. This will affect the final consensus low-rank representation learned from all views. 2CMV disregards the close distance information if it happens on one or some views but not on all views, and aims to learn the optimal manifold encoding all common distance information.

The second approach supposes the existence of a consensus manifold in multi-view data and assumes that the consensus low-rank representation should be smooth with the consensus manifold [120]. However, it is a non-trivial task to learn the accurate consensus manifold as well as it can be a very strict constraint to require the consensus low-rank matrix to be smooth with the consensus manifold in the new low-dimensional space. 2CMV does not require the consensus matrix to respect the consensus manifold, however, it learns the optimal manifold and allows the low-rank matrix on each view to be guided by the optimal manifold.

5.6.4 Enhancing the Complementary Information of the Learning Process

In this section, we explain how 2CMV incorporates the *complementary enhancing term* into the objective function based on the concept of orthogonality that will be applied on two components, H_* and H_{com}, denoted as CE(H_*, H_{com}).

Suppose we have two factor matrices H_* and H_{com} as illustrated in Figure 5.7. Let h_*^i express the latent features of data sample ith encoded in the consensus latent feature matrix. Let h_{com}^i be the latent features of data sample ith encoded in the complementary latent feature matrix. We want to encourage the learning process to learn as much distinctive information as possible for the two components and thus more complementary information can be learned for the data sample ith. For this purpose of enhancing the independence of the two vectors, these two vectors should be orthogonal in the low-order space and this fact will be equivalent to minimizing the inner product of the two vectors. Generally, minimizing $Tr(H_* H_{com}^T)$ will lead

to ($H_* \perp H_{com}$) when the most complementary information for the new low-rank representations has been promoted. We design the complementary enhancing term as,

$$\mathrm{CE}(H_*, H_{com}) = \min Tr(H_* H_{com}^T) \tag{5.39}$$

Since the consensus component is learned from the shared mode in CMF, H_* is believed to embed the most common information among views. Our aim of adding the constraint is to allow the complementary component to carry the maximum feasible distinct information from different views.

5.6.5 The Final Objective Function: 2CMV

Combining Eqs. (5.33), (5.37) and (5.39), we arrive at the final objective function as below,

$$\begin{aligned} J = \min & \sum_{v=1}^{n_v} \left(\|X_v - [H_* \ H_v] W_v^T\|_F^2 + \lambda Tr(H_v^T L_{opt} H_v) \right) + \delta \mathrm{CE}(H_*, H_{com}), \\ \text{s.t. } & H_{final} = [\alpha H_* \ (1-\alpha) H_{com}], 0 < \alpha < 1, \\ & H_* \geq 0, H_v \geq 0, W_v \geq 0, H_{com} = \sum_{v=1}^{n_v} \frac{H_v}{n_v} \end{aligned} \tag{5.40}$$

For the purpose of ensuring an even distribution on all clusters and avoiding over-fitting, we apply $l2$-normalization on all factors of the proposed model including H_*, H_{com} and W_v. The first term in Eq. (5.40) is to learn the two components. The graph regularization term corresponding to the novel optimal manifold in the second term will ensure maintaining the most important geometric information in multi-view data while enhancing both the compatible and complementary information with the assistance of the third term, leading to the complete method of 2CMV for learning both consensus and complementary information designed for multi-view data. Algorithm 4. summarises the main steps of 2CMV.

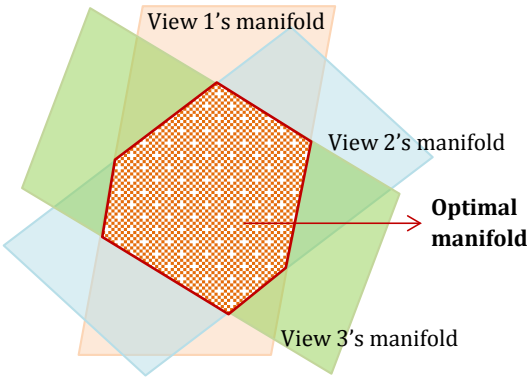


Figure 5.6: The optimal manifold learned from a two-view dataset.

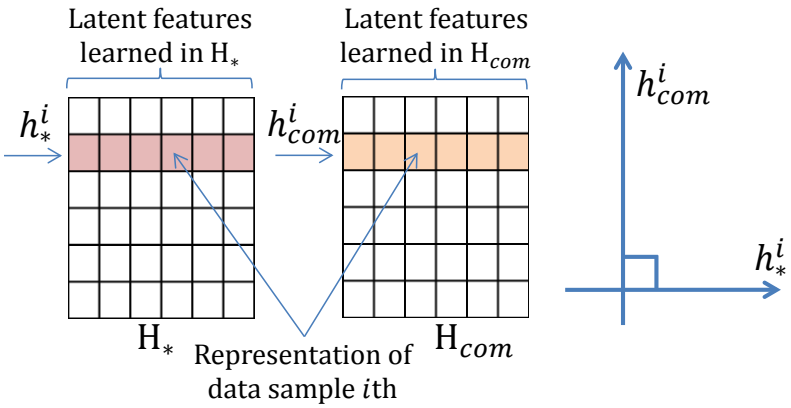


Figure 5.7: Illustration of consensus and complementary low-rank representations.

5.6.6 Algorithmic solution

5.6.6.1 Solving H_*

The objective function J in Eq. (5.40) with respect to H_* can be written as,

$$J_{H_*} = \min \sum_{v=1}^{n_v} \|X_v - H_* W_v^T\|_F^2 + \delta Tr(H_* H_{com}^T) \quad \text{s.t. } H_* \geq 0 \tag{5.41}$$

By taking the first derivative of J_{H_*} on H_* we have,

$$\frac{\partial J_{H_*}}{\partial H_*} = -2 \sum_{v=1}^{n_v} X_v W_v + 2 \sum_{v=1}^{n_v} H_* W_v^T W_v + \delta H_{com} \tag{5.42}$$

Table 5.5: Characteristic of the datasets

Properties	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
# Samples	165	2,000	2,908	2,708	12,100	2,400
# Feature View 1	4,096	76	20	1,433	7,000	6,000
# Feature View 2	3,301	47	59	5,429	4,000	6,000
# Feature View 3	6,750	-	40	-	3,000	-
# Feature View 4	-	-	-	-	5,000	-
# Classes/# Views	15/3	10/2	10/3	7/2	6/4	10/2

Table 5.6: NMI for each dataset and method

Methods	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**
NaiveMV	57.93	72.52	27.21	33.17	27.48	51.02
MultiNMF	62.24	68.28	41.14	30.48	24.61	51.56
MVCMF	62.08	77.84	44.95	36.88	29.75	54.14
DiNMF	59.83	75.51	39.37	23.35	31.43	56.88
LP-DiNMF	63.89	85.17	40.00	26.63	33.82	59.01
MMNMF	64.73	85.91	40.38	34.22	**37.70**	61.57
2CMV-CE	60.55	89.76	43.45	32.87	30.47	58.42
2CMV-OM	67.87	92.15	52.54	37.14	34.34	**63.50**
2CMV	**72.32**	**92.26**	**56.37**	**39.01**	35.61	63.01

Following the similar condition as in [29] for non-negativity of H_*, we have the update rule for H_*,

$$(H_*)_{ij} = (H_*)_{ij}\Big[\frac{(\sum_{v=1}^{n_v} X_v W_v)_{ij}}{(\sum_{v=1}^{n_v} H_* W_v^T W_v + \frac{\delta}{2} H_{com})_{ij}}\Big]^{1/2} \tag{5.43}$$

5.6.6.2 Solving H_v

The last term in the objective function Eq. (5.40) can be re-written as,

$$\delta \text{CE}(H_*, H_{com}) \Leftrightarrow \delta Tr(H_*(\sum_{v=1}^{n_v} \frac{H_v^T}{n_v})) \tag{5.44}$$

To find the update rule for H_v, v and thus W_v will be fixed. The objective

Table 5.7: Accuracy for each dataset and method

Methods	D1	D2	D3	D4	D5	D6
NaiveMV	55.15	75.90	35.39	55.24	39.35	51.54
MultiNMF	58.18	68.70	48.59	46.16	41.07	50.08
MVCMF	58.18	85.60	52.20	52.22	46.70	52.71
DiNMF	57.58	84.45	43.43	44.22	42.97	54.33
LP-DiNMF	59.39	91.30	46.11	50.48	44.84	60.21
MMNMF	61.82	91.41	41.33	55.02	52.86	63.92
2CMV-CE	55.76	94.05	51.44	**58.53**	43.76	58.67
2CMV-OM	66.67	96.55	58.91	50.63	44.55	**65.58**
2CMV	**70.30**	**96.65**	**65.75**	54.36	**58.10**	**65.58**

Algorithm 4: 2CMV

Input : Multi-view data matrices and $\{X_v\}_{v=1..n_v}$; affinity matrices $\{A_v\}$ of all views; parameters $\alpha, \beta, \lambda, \delta$ and k with non-negative values as discussed.

Output: Cluster indicator matrix of the multi-view data.

Initialize non-negative matrices H_v by K-means, H_* and H_{com} are initialized as a linear combination of $\{H_v\}_{v=1..n_v}$

Calculate the optimal manifold L_{opt} as in Eq. (5.34)

repeat

- Update H_* as in Eq. (5.43)
- Normalize H_*
- Update H_v as in Eq. (5.47) for all $v = 1..n_v$
- Normalize H_v
- Update W_v as in Eq. (5.50) for all $v = 1..n_v$
- Normalize W_v

until *converges*;

Calculate H_{final} as given in Eq. (5.40) and transform H_{final} into cluster indicator matrix by K-means

function J with respect to H_v is as follows,

$$J_{H_v} = \|X_v - H_v W_v^T\|_F^2 + \lambda Tr(H_v^T L_{opt} H_v) + \delta Tr(H_* \frac{H_v^T}{n_v}), \text{ s.t. } H_v \geq 0 \tag{5.45}$$

By introducing the Lagrangian multiplier matrix Ψ and setting the first derivative to 0 we have,

$$\Psi = -2X_v W_v + 2H_v W_v^T W_v + 2\lambda L_{opt} H_v + \frac{\delta}{n_v} H_* \tag{5.46}$$

Since the Karush-Kuhn-Tucker (KKT) condition [18] for the non-negative constraint on H_v gives $(\Psi)_{ij}(H_v)_{ij} = 0$, we have the following updating rule for H_v,

$$(H_v)_{ij} = (H_v)_{ij}\Big[\frac{(X_v W_v + \lambda L_{opt}^{-} H_v)_{ij}}{(H_v W_v^T W_v + \lambda L_{opt}^{+} H_v + \frac{\delta}{2n_v} H_*)_{ij}}\Big]^{1/2} \tag{5.47}$$

$L_{opt}^{+/-} = [|(L_{opt})_{ij}| \pm (L_{opt})_{ij}]/2$ [30].

5.6.6.3 Solving W_v

When fixing v and thus fixing H_v, we have the objective function in Eq. (5.40) with respect to W_v can be written as,

$$J_{W_v} = \min \|X_v - [H_* \; H_v] W_v^T\|_F^2, \text{ s.t. } W_v \geq 0 \tag{5.48}$$

Taking the first derivative of J_{W_v} on W_v we have,

$$\frac{\partial J_{W_v}}{\partial W_v} = -2X_v^T H_* + 2H_* H_*^T W_v e - 2X_v^T H_v + 2H_v H_v^T W_v \tag{5.49}$$

Following the similar condition in [29] for non-negativity of W_v, we have the update rule for W_v

$$(W_v)_{ij} = (W_v)_{ij}\Big[\frac{(X_v^T H_* + X_v^T H_v)_{ij}}{(W_v H_*^T H_* + W_v H_v^T H_v)_{ij}}\Big]^{1/2} \tag{5.50}$$

5.6.7 Time complexity

The main complexity of 2CMV lies in calculating L_{opt}, H_*, $\{H_v\}$ and $\{W_v\}$, $\forall v = 1..n_v$. Calculating L_{opt} is dominated by calculating A_{opt} which is computed from $\{A_v\}_{v=1..n_v}$. Constructing A_v requires $O(n^2 k)$ to construct k non-zero elements for each row of A_v using the kNN graph. Thus A_{opt} has complexity of $O(n_v n^2 k)$. Calculating $\{W_v\}$ requires $O(n_v n^2 r)$ where n_v is the number of views; feature sizes m_v and sample size n are supposed to be similar. Calculation of H_*, $\{H_v\}$ requires about the same time complexity to the cost for $\{W_v\}$ since the proposed comple-

mentary enhancing term that relied on minimizing the trace of the product of two matrices that does not require any additional cost. Therefore, the computational complexity of 2CMV is quadratic where construction of the optimal manifold and complementary enhancing will not incur extra cost.

5.7 Experiment Analysis

An extensive experiment has been conducted to check the performance of 2CMV on several real-world datasets in comparison to the state-of-the-art multi-view clustering methods. We utilize two popular measurement criteria in clustering, i.e., accuracy, the percentage of correctly obtained labels and Normalized Mutual Information (NMI) [119]. Average results are reported after 5-fold runs of each experiment.

5.7.1 Datasets

We used the well-known and wide-range of text and image multi-view datasets to evaluate the performance. The task is to find subgroups in the data that share the highest commonality. The number of clusters on each dataset is set to the actual number of classes, as shown in Table 5.5.

Yale (D1)[4]: A popular image dataset which consists of 165 images of 15 subjects corresponding to faces in raw pixel. Each subject is expressed by 11 images with different conditions such as facial expressions, illuminations, with/without glasses, lighting conditions, etc. We chose three different views, i.e., intensity Historgram, Local Binary Pattern (LBP) and Gabor features [111].

UCI Handwritten Digits (D2)[5]: It includes 2000 examples. Two views selected are Fourier and Zernike.

Caltech101 image data set (D3)[6]: a subset was extracted from the most popular

[4] http://vision.ucsd.edu/content/yale-face-database
[5] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[6] http://www.vision.caltech.edu/Image-Datasets/Caltech101/

multi feature dataset. We extract GIST, LBP and HOG descriptors as the three views and select the first 10 classes for the dataset.

Cora (D4)[7]: A scientific publication dataset with documents represented via two views, content and cites.

Multilingual Reuters (MLR) (D5): a subset was extracted from the most popular multi-view dataset of Reuters RCV1/ RCV2 Multilingual [7] with English, French, Italian and German terms representing each view.

Reuters-21578 (D6): a subset was generated to contain 10 largest classes in the dataset. Apart from the first view where documents are represented via terms, we used external knowledge, i.e., Wikipedia, to represent the second view of components (following steps as in [48, 72]).

5.7.2 Benchmark Methods

To our best of knowledge, all relevant state-of-art methods were used for comparison.

1. NaiveMV: an implementation of using NMF to factorize on all view data and conduct K-means on the consensus low-rank representation learn by linearly combining all representations of all views.

2. MultiNMF [56]: the low-dimensional representations from all views are simultaneously learned before seeking the consensus low-dimensional matrix.

3. MMNMF [120]: the consensus coefficient matrix is learned by linearly combining low-rank coefficient matrices of all views and forced to follow the linear consensus manifold. For easy comparison with other methods, we use Euclidean distance instead of Kullback-Leibler in the loss function.

4. DiNMF [103]: A NMF-based multi-view method with the constraint on each pair of low-rank representations of two views to enhance the diversity among multi views.

[7] http://www.cs.umd.edu/Esen/lbc-proj/data/cora.tgz

5. LP-DiNMF [103]: The method is based on DiNMF with the incorporation of the graph regularization to ensure the local geometry structure on each view.

6. MVCMF [71]: Coupled matrix based multi-view method with the Multiplicative Update Rule instead of Greedy Coordinate Descent to have a fair comparison between benchmarking methods.

Importantly, to investigate the value of the proposed optimal manifold learning and the complementary enhancing term, we implemented two versions of 2CMV, called as **2CMV-OM**, i.e., 2CMV with optimal manifold learning only (i.e., without the complementary enhancing term) and **2CMV-CE**, i.e., 2CMV with the complementary enhancing term only (i.e. without the optimal manifold learning).

5.7.3 Clustering Results

As shown in Table 5.6 and Table 5.7, 2CMV clearly outperforms other benchmarking methods demonstrated by a higher NMI and Accuracy values on all datasets except on D5 where MMNMF beats 2CMV on NMI but trades-off by a much lower accuracy. All six benchmarking methods are matrix factorization based, while the first four methods are without manifold learning technique and thus fail to respect the geometric shape of data. The last two methods rely on learning and preserving the geometric shape of the original data when projecting to the lower dimensional space. With the aid of manifold learning, the latter two benchmarking methods, LP-DiNMF and MMNMF, achieve higher results as compared to other benchmarks. MMNMF produces higher results than LP-DiNMF that proves the importance of learning the consensus manifold of multi-view data instead of simply learning and preserving the manifold of data on each view as in LP-DiNMF. MMNMF also performs the best among benchmarking methods however can be easily beaten by 2CMV-OM where we use the proposed optimal manifold only which can prove the effectiveness of the proposed optimal manifold as compared to the consensus manifold learned in MMNMF. More importantly, MMNMF is far beaten by 2CMV when both the

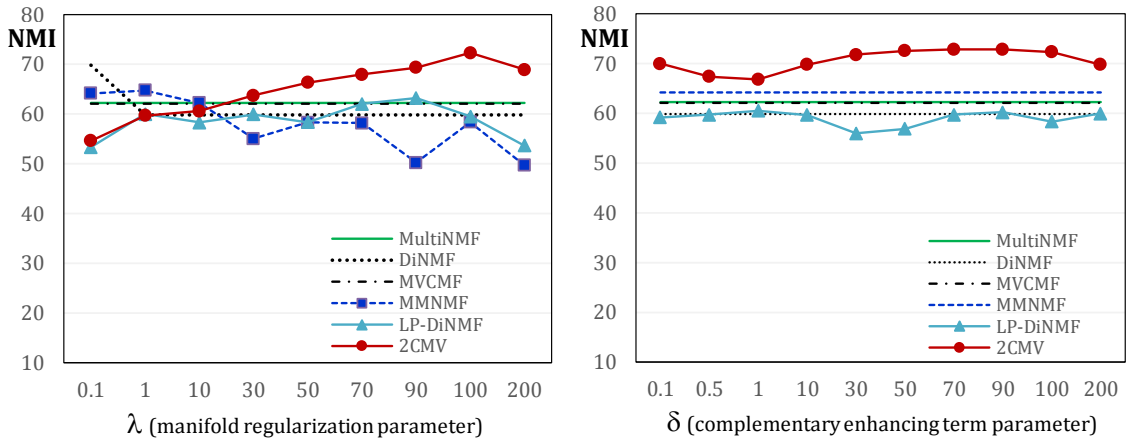


Figure 5.8: NMI changes with the alterations of λ and δ on Yale (D1)

optimal manifold and the complementary enhancing term are activated. This shows the effectiveness of 2CMV with the aid of both optimal manifold learning and complementary enhancing term.

We also observe that among the first four benchmarking methods, MVCMF produces a fairly good results as compared to other, because it learns the consensus low-rank representation directly and naturally from all view data, thus can embed the best consensus representation and can exploit the associative relationship between different view data. Furthermore, 2CMV-CE, the incomplete method when we use the complementary enhancing term only, cannot beat LP-DiNMF and MMNMF, but it can easily beat DiNMF and other first four benchmarking methods most of the time. This proves the role of enhancing the complementary between the consensus and complementary components for multi-view data as well as it explains the superiority of 2CMV with an integrated NMF and CMF framework with emphasis on learning both the consensus and complementary information effectively.

5.7.4 Parameter setting

Except for the first three methods, NaiveMV, MultiNMF and MVCMF are free with parameters, the other methods need a selection for their corresponding parameters.

We set those values autonomously by searching from the grid with a range of values. To enable a fair comparison between 2CMV and other methods, we set the same range of values for the common parameters in this paper. All methods rely on manifold learning for their purposeful solutions such as MMNMF, LP-DiNMF and 2CMV, hence they need the manifold regularization parameter λ and the neighbourhood size k tunning. Similar to prior research [43, 48], the value for λ is selected by searching from the grid $\{0.1, 0.5, 1, 10, 50, 100, 200\}$ for all methods. For the neighbourhood size k, previous works [21, 74] have pointed out that the value for k should be chosen to be approximating $log(n)$, n is the number of samples. We check on a wide range of value for k, the highest result is achieved when $k = 5$ or $k = 10$ depending on dataset, this justifies that selection of k for our model is not different to other kNN graph-based methods. The results reported are the highest result when these parameters selecting in their ranges on all methods. For example, on dataset D1, 2CMV achieves its highest results when $\lambda = 100$ and $k = 10$, while MMNMF hits its peak when $\lambda = 1$ and $k = 5$.

The complementary enhancing term parameter δ proposed in 2CMV is set depending on the level of the compatible and complementary information exist among multiple views of the dataset. We set the value for δ by searching from the grid $\{0.1, 0.5, 1, 10, 50, 100, 200\}$. Two benchmarking methods, DiNMF and LP-DiNMF, need an extra diversity parameter, we let this diversity parameter to search in the same wide grid and report the highest result.

Apart from the above common parameters, 2CMV has α, the two-component model parameter, to generate the final low-rank representation H_{final}. Note that learning of H_* and H_{com}, as given in the corresponding update rule (Eqs. (5.43), (5.47)), does not depend on the parameter's selection. α is used at the final stage when two matrices H_* and H_{com} have been learned to combine these two components to achieve the best H_{final}. It is natural to use $\alpha = 0.5$ when the two components can contribute equally to decide H_{final}. Alternatively, the value for α can

be autonomously picked from the grid $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. 2CMV produced the highest result when α is selected from $\{0.3, 0.4, 0.5\}$ for all datasets, which justifies the need of using both components as proposed in the method.

Figure 5.8 illustrates the influences of different values for manifold parameter λ and complementary enhancing term parameter δ on the clustering performance on dataset D1. It shows that 2CMV easily beats other benchmarking methods on a wide range of parameter selections.

5.8 Conclusion

The paper has proposed a novel integrated factorization based model, 2CMV, to learn both the consensus and complementary information from all view data. Moreover, 2CMV incorporates the process of learning the optimal manifold as well as it adds a complementary enhancing term that lead to a superior multi-view learning method. Empirical analysis shows that this method outperforms the state-of-the-art clustering methods.

Chapter 6

Conclusion

The rapid growth of computational technology has resulted in data exhibiting multiple aspects with the high inter-relatedness between them. The problem of learning on multi-aspect data has gained much interest and increasingly becoming important, not only for the clustering task but for many other tasks such as classification and pattern recognition. Though there have been a wide range of methods developed for clustering on the multi-aspect data, due to the complex nature of the data, there exist many challenges, as highlighted in Chapter 2. This research project has contributed to the problem of multi-aspect data learning by providing advanced methods for learning accurate clusters.

This chapter will summarise the research contributions and main findings as well as report the future works in the end of the chapter.

6.1 Research Contributions

In order to fill the gaps highlighted in Chapter 2, the thesis has provided a comprehensive survey paper and four novel factorization-based methods. The survey paper focuses on the group of methods that use NMF and manifold learning for data understanding. The methods can be applied to a wide range of dataset types ranging

from traditional one-aspect data to multi-aspect data. The main contributions are summarised as follows.

- A comprehensive review on a wide range of the important class of multi-aspect data clustering methods, focusing on the group of methods based on an NMF framework utilising manifold learning algorithms, has been conducted. Based on the properties of the methods, different categories of methods have been surveyed; as well, many instances of multi-view data have been investigated. An extensive experiment on several datasets has been conducted; as well, many open problems have been raised so a higher clustering performance can be achieved. This survey paper has (1) provided a general view of how to understand multi-view data via applying NMF and learning manifold; (2) included an experimental study on well-known multi-view datasets to investigate the effectiveness of the methods on datasets and (3) discussed the challenging problems that can be addressed in future research. This is the first survey paper to summarize a class of clustering methods combining NMF and manifold learning to effectively understand the challenging multi-view data problem.

- Most manifold learning based methods, which rely on the local geometric structure for a better clustering solution, fail to utilize the global geometric structure for data. This may lead to an unrealistic representation of data. The thesis has presented ARASP (Learning Association Relationship and Accurate Geometric Structure or MTRD) where the contributions are threefold. First, the method works well on MTRD data to learn the meaningful clusters that respect the accurate geometric structure of the original data. Second, ARASP is the first method proposed to incorporate the association relationship between clusters of different object types into the learning process to help obtain general correlation in the data. The inclusion of the association relationship helps the learning process to be faster converged. Third, the new fashion of

constructing affinity matrix, where both important close and far distances have been considered, can be a promising technique to be applied in a dimensionality reduction problem. The method outperforms other benchmarking methods on MTRD, evidenced by its effectiveness.

- Most MTRD clustering methods formulate the intra-relationship using manifold learning algorithms, and ignore the manifold generated by the inter relationship between objects of different types, and this may lead to a less diverse and less complete manifold learned. In DiMMA (Diverse Manifold for Multi-Aspect data), the inter-manifold concept is proposed and integrated with the normal intra-manifold to form a complete and diverse manifold for data representation learning. This manifold is then steadily preserved during the learning process to ensure the new low rank representation is smooth with the original manifold, and therefore can produce a purposeful cluster solution. The DiMMA method is designed under the context of the general multi-aspect data in the most genetic way, where the method can perfectly work on both MTRD and multi-view data. This work also provides a comprehensive analysis of vital factors of the clustering task such as ranking, relatedness between data and geometric structure of data.

- Most NMF-based multi-view learning methods learn the consensus low-rank matrix separately and independently via two separate steps: learning the low-rank matrix for each view and learning the consensus low-rank matrix from all views. There exists no method that can simultaneously and jointly learn the consensus coefficient matrix for multi-view data in order to ensure maintaining the consensus and complementary information for multi-view data. MVCMF (Multi-view Coupled Matrix Factorization), is able to fill the gap and can provide the unified framework for multi-view data clustering. In MVCMF, the coupled matrix factorization is deployed in multi-view data representa-

tion learning in an effective manner to simultaneously and jointly learn the consensus common low-rank representation from all data views. The advantages of the coupled matrix factorization process are that it not only naturally produces the meaningful representation for multi-view data, it also saves preforming time for the learning process because the number of factor matrices have been reduced to half. Besides, the inclusion of affinity information between data samples in the factorizing process has helped in preserving the closeness information between samples when data is projected from a high to a lower dimensional space. The MVCMF method can be considered as a fundamental method where other regularizations and constraints can be applied in order to achieve the better performance.

- Multi-view data is beneficial to the learning process by providing both the consensus and complementary information. The crucial principles of learning multi-view data is ensuring the learning process to effectively utilize the consensus and complementary information for the data. While most existing methods neglect or fail to adhere to these two principles, 2CMV (Learning the Consensus and Complementary information for Multi-View data) is proposed to fill this gap. Both consensus and complementary information have been comprehensively exploited in 2CMV where a new integrated framework using both NMF and CMF has been used; a novel optimal manifold embedding the most consensus manifold of multiple views has been proposed and a novel complementary enhancing term has been imposed on the consensus and complementary components learned from the model. 2CMV is also the first method to propose the optimal manifold concept for multi-view data, in an effort to ensure both the consensus and complementary information will be learned adequately for an insightful clustering solution.

The proposed methods have contributed to the problem of multi-aspect data

Method	Dataset Type	Framework	Manifold Learning	Constraint
ARASP*	MTRD and Multi-view	NMF	Learn close and far distances	Normalized Cut-Type; Non-negativity; l1-norm
DiMMA	MTRD and Multi-view	NMF	Learn intra and inter distances	Non-negative; l1-norm
MVCMF	Multi-view	CMF	Factorize similarity matrix	Non-negativity
2CMV	Multi-view	CMF and NMF	Learn optimal manifold	Non-negativity; l2-norm; Complementary enhancing

* This method also learns the association relationship between clusters of different object types and is used as an application to the problem of Community Discovery.

Table 6.1: A summary of proposed methods' characteristics

learning by enhancing the clustering performance to achieve the accurate clustering results, however, they work distinctly. The first two methods (ARASP and DiMMA) utilize an NMF framework and aim to obtain good representation relying on learning the accurate manifold. They can be straightforward applied to the task of dimensionality reduction. The later two methods (MVCMF and 2CMV) aim to utilize a new framework that can be the best fit for a multi-view learning problem. The optimal manifold constructed in 2CMV can be used in other clustering approaches based on dimensionality reduction for multi-view data. In addition, the approach of using CMF or the integrated framework using both NMF and CMF can be extended to many other problems.

It should be noted that, (1) all of the proposed methods, though designed for multi-aspect data, can be easily applied on traditional one type or one view data and (2) while MVCMF and 2CMV are designed for a multi-view dataset only, the other two methods can work well on both MTRD and multi-view data. A summary on the characteristics of the four proposed methods is given in Table 6.1.

An additional contribution made in this thesis is the application of one of ARASP

in an application domain. In this thesis, it is the first time an application of the multi-aspect data clustering method has been deployed on the problem of community discovery and brought the promising results. This shows the importance of multi-aspect data learning in real-world problems.

6.2 Answers to Research Questions and Findings

6.2.1 Response to Question 1 regarding to Data and Findings

In response to the first research question about the distinct characteristics of MTRD and multi-view data as compared to traditional data that can affect the performance of clustering, the answers are given as below.

The most important characteristics of multi-aspect data, distinguished from the traditional data, are the interrelatedness between data that requires the clustering process to be simultaneously conducted on multiple available aspects of data. When clustering on multi-aspect data, it is important to include all available information to produce a meaningful solution. Each dataset type may have a distinct concern, for instance, the concern of clustering MTRD is to make use of all possible relationships to simultaneously cluster all object types, such that all interrelatedness between different type objects has been utilized. The concern of clustering on multi-view data is to make use all data views learn the consensus common data representation from all views that ensure embedding the compatible and complementary information. The thesis has found that different requirements of different dataset types may affect the working manner of learning methods.

6.2.2 Response to Question 2 regarding to Methods and Findings

In response to the second question about the groups of methods that should be selected to effectively work on multi-aspect data and the associated findings of the methods applied on multi-aspect data, following are the answers for the detailed questions.

6.2.2.1 For the Question 2.a of why NMF and manifold learning framework is applied in multi-aspect data

Due to exposing the sparseness and high-dimension, dimensionality reduction technique-based methods can be the prominent selection. The factorization-based methods with the aid of manifold learning are selected in this research project due to the ability of NMF to act as feature transformation for the clustering task and to be easily extended to multi-aspect data. In this way, the NMF framework is able to return the dense part-based representation.

6.2.2.2 For the Question 2.b of how NMF and manifold learning framework is deployed in multi-aspect data

In MTRD, due to the inter-relatedness between objects of different object types in the dataset, the factorization process on all input data matrices should be done simultaneously in order to make use of the inter-relatedness among data objects and to produce the dense part-based representations for all object types before conducting clustering. Furthermore, the thesis found that the factorization process should be designed such that it can be able to include all possible relationship types available in the MTRD.

In multi-view data, the main task is to learn the consensus representation embedded in multiple views. This consensus low-rank representation should be used as

the input for the clustering task where the compatible and complementary information from all views has been embedded. Due to the hard principle of ensuring both the consensus and complementary information in clustering, the thesis found that trivial extensions of the NMF framework to multi-view data may not deliver good performance. Therefore, the coupled matrix factorization framework or an integration of both NMF and CMF can be the most effective framework for multi-view data learning, where the most important consensus and complementary information embedded in the original multi-view data can be learned adequately.

6.2.2.3 For the Question 2.c and Question 2.d of how will an NMF framework and manifold learning-based method achieve the highest outcome and their associate challenges

The thesis has found two main challenges of incorporating manifold learning in NMF framework on multi-aspect data with regard to learning the accurate and useful manifold for multi-aspect data and choosing the neighbourhood size when constructing the affinity matrix for learning the data manifold.

Firstly, it has been already showed by the prior research that manifold learning is highly beneficial to the NMF learning process. The research in this thesis found that the more meaningful the manifold can be learned, the more meaningful and accurate clusters can be achieved. This research investigated the three means of building meaningful manifolds.

1) The accurate manifold

This is achieved by incorporating both close and far distances to learn both the local and global geometric structure of data. Though, most manifold learning-based clustering methods construct manifold that utilizes the local geometric structure of data since the local geometric shape is believed to be more beneficial to clustering task. However, this research has shown that not only the local geometric shape of data, but the global geometric shape also plays an important role in clustering

problem. It will be the premise for obtaining more accurate and informative clusters.

2) The diverse and complete manifold

The thesis discovered that the manifold learning technique can be exploited on inter-type relationships in MTRD data to construct the inter-manifold, thus can produce the diverse and informative manifold for data by combining both intra- and inter-manifold. The diverse manifold, as constructed in DiMMA, can help the clustering process to be conducted more effectively.

3) The optimal manifold

The concept of consensus manifold in multi-view data has been used in previous work, however, it is a non-trivial task to learn the correct consensus manifold embedded in multiple views. The thesis found that there exists an optimal manifold, which is the most consensus manifold embedded in all view data. This optimal manifold will contain the most common closeness information of data points from all views; therefore constraining the low-rank representations learned from different views to be smooth with this manifold is essential in guaranteeing that the new representation is satisfactory.

Secondly, since most manifold-based clustering methods mainly take advantage of a local geometric shape, the manifold learning algorithms are based on constructing a kNN graph to model the neighbourhood distances between points. However, the difficulty of using a kNN graph is that the neighbourhood size k is unknown beforehand. The thesis found that the dependence in choosing k can be reduced by using a novel fashion in constructing the affinity matrix of incorporating both close and far distances, as in the proposed ARASP method.

6.2.3 Response to Question 3 regarding to Relationships and Findings

In response to the question related to relationship types used in multi-aspect data, the findings are as follows.

The concept of relationship is not the focus in multi-view data, however it is the main challenge and the focus for the MTRD. Most existing MTRD clustering methods use intra-type and inter-type relationships for their meaningful solutions, however, an emerging question will be how to learn accurate intra-relationship and effectively utilize other relationship types.

Since the intra-relationship type is constructed based on constructing the kNN graph that heavily depends on choosing the neighbourhood size k. To learn the accurate intra-relationship, it is essential to look at data in a different perspective such as what types of intra information can be utilized. The research found that not only the close distance, but the far distance also plays an important role for an accurate and meaningful representation.

Apart from intra-type relationships in MTRD, to successfully take advantage of an inter-type relationship, it is vital to make use of all types of information it may carry. For example, not only using inter-type relationship information encoded in input matrices for the learning process, but looking at its geometric structure could also bring much more useful information for the learning process.

To achieve the higher clustering performance, other relationship types, i.e., association relationships between clusters, need to be taken into account in order to utilize as much information as possible and produce higher clustering performance. The thesis has found that the inclusion of association relationship into the learning process can boost the learning process to be faster converged and bring out more meaningful representation.

In the context of multi-view data, though the relationship is not the main focus,

exploiting the available relationship in the data is believed to bring more useful solution; the use of coupled matrix factorization that explores the associative relationship among all views is evidence of this. This will assert the assumption that the more information used, the more accurate clustering solution will be obtained.

6.2.4 Response to Question 4 regarding to Constraints and Findings

The following findings are interpreted from answering question 4, which is related to using constraints in multi-aspect data clustering methods.

The research project has used and investigated many types of constraints. The most common constraints found to be useful in NMF framework are non-negative and $l1$-normalization. The non-negative constraint applied on factor matrices of the NMF factorizing process enables learning the part-based representation or doing the feature transformation that will be helpful for the clustering process. The l1-norm applied on factor matrices makes the factor matrices more meaningful, since it comes close to the possibility of how a data point belongs to a cluster.

The orthogonal constraint applied on the NMF framework, together with the non-negative constraint, will turn the factor matrices into indicator matrices, as has been noted in the previous work [29]. The orthogonal constraint is useful in helping return a unique and interpretable solution, however this constraint is found to be too strict and very slow in convergence.

A variant of the orthogonal constraint is the Normalized-Cut Type constraint [93], the well-known constraint in spectral based clustering methods. When it is used in NMF framework, it can work as a pseudo-orthogonal constraint. Thus, it can help the learning process to be converged faster and can result in the factor matrices being as unique as possible. Furthermore, with the help of this constraint, the trade-off matrix becomes the association relationship encoding the interactions

between clusters of different object types as reported in the ARASP method.

6.2.5 Response to Question 5 regarding to Application and Findings

In order to answer the question about the possible applications of the MTRD and multi-view clustering methods, the following findings can be reported. First, the multi-aspect data clustering methods, applied on multi-aspect data, are getting more and more popular nowadays due to the ubiquity and complexity of the modern data generated.

Second, it has been proved in this thesis that while it is infeasible or inefficient to use a traditional one-aspect data clustering method for multi-aspect data clustering, all MTRD and multi-view clustering methods proposed in this thesis can be easily deployed on traditional one type data, co-clustering (bi-type data) or one-view data.

Third, recent works have reported that the problem of clustering comes close to the problem of community detection and there have been clustering methods successfully applied on community discovery. The thesis has made the investigation of applying an MTRD clustering approach on community discovering and asserted that this is a promising field for the developed multi-aspect learning methods.

6.3 Future Works

The thesis aims to develop clustering methods for multi-aspect data including MTRD and multi-view data. Various improvements can directly be applied to the proposed methods. The extended directions can be carried out, presented as below.

6.3.1 Future works related to proposed methods

- Many works in the literature have shown that manifold learning is highly related to data distribution, so a possible work that can be extended from the

first method (ARASP), related to constructing the accurate and meaningful manifold for data representation, is to make use of learning the distribution of data on original space and preserve them in low-order space. The success of this work will eliminate the dependence of parameter k in manifold learning based techniques and will learn the more accurate and more informative manifold for group learning.

- The second method (DiMMA) aims to design a comprehensive framework for multi-aspect data clustering. The symmetric affinity matrix can also be decomposed together with the inter-type relationship matrices to obtain the low-rank factor matrices from both intra- and inter-type relationships. Moreover, three terms to improve performance of a clustering method, i.e., intra-manifold learning, inter-manifold learning and subspace learning, have been mentioned in the work, so an investigation of how the characteristic of a dataset relates to choosing the right parameters for the three terms in order to achieve the highest performance for clustering can be a promising direction for future work.

- From the third method of proposing MVCMF where a coupled matrix has been used, there are some directions can be made. First, a close but more general framework that is a tensor model can also be used for multi-aspect data. We will further investigate to extend the proposed framework in tensor factorization to handle higher order multi-aspect clustering. Second, an investigation on the insight of different updating schemes i.e., HAL, FHALs, GCD, MUR, as well as how they affect the final results can be conducted. Third, it is also interesting to further quantify the importance of the usage of the similarity matrix as a extra view as in MVCMF.

- The whole thesis has used *tf-idf* weight for text dataset representing. However, a recent emerging trend of using word embedding to represent text datasets may be worthy to examine in the context of the clustering multi-aspect data

considered in this thesis.

- In this thesis, we aim to treat all aspect data equally so we do not turn the parameters on for different aspects or different relationships. However, other ways of treating different aspects or combining different relationships, for example, a weighted sum, can be considered in future work.

6.3.2 Future works related to extended applications

ARASP has been applied in community discovery and shown a promising outcome. Other methods can also be deployed in community discovery with some adjustments. Furthermore, the NMF framework to learn the low-rank representation has also been proved as an effective framework for collaborative filtering. Combining NMF and manifold technique as used in this thesis can be extended for the collaborative filtering, utilising the context of multi-aspect data.

6.3.3 Future works related to extended direction

First, a closely related research topic to the clustering task dealt in this thesis is the dimensionality reduction task. All techniques in the thesis can be adjusted and designed for dimensionality reduction and used in other research fields such as pattern recognition, computer vision, bio-informatics.

Second, Sammon [45] is an algorithm to map data from high dimensionality to a lower dimensional space. It is a non-linear subspace learning-based approach and is able to maintain the inter-pattern distances between points when mapping. In future work, it should be investigated whether Sammon can be used in multi-aspect data learning.

Third, a deep learning technique has shown effectiveness in supervised learning. Investigating how the deep learning technique can be deployed in multi-aspect data learning is also a promising direction that requires attention.

Bibliography

[1] E. Acar, R. Bro, and A. K. Smilde. Data fusion in metabolomics using coupled matrix and tensor factorizations. 103:1602, 09 2015. 155, 176, 179

[2] E. Acar, G. Gurdeniz, M. A. Rasmussen, D. Rago, L. O. Dragsted, and R. Bro. Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, 3(3):22–43, 2012. 155

[3] C. C. Aggarwal and C. K. Reddy. *Data Clustering: Algorithms and Applications*. Chapman Hall CRC, 1st edition, 2013. 116

[4] C. C. Aggarwal and C. Zhai. *A Survey of Text Clustering Algorithms*, pages 77–128. Springer US, Boston, MA, 2012. 28, 51

[5] M. Akbari and T.-S. Chua. Leveraging behavioral factorization and prior knowledge for community discovery and profiling. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 71–79, New York, NY, USA, 2017. ACM. 10

[6] M. Akbari and T.-S. Chua. Leveraging behavioral factorization and prior knowledge for community discovery and profiling. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 71–79. ACM, 2017. 93, 95

[7] M. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In Y. Bengio,

D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 28–36. Curran Associates, Inc., 2009. 135, 191

[8] D. AParrochia and P. Neuville. *Towards a General Theory of Classifications*. Springer Basel, 2013. 29

[9] F. R. Bach and M. I. Jordan. Learning spectral clustering. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 305–312. MIT Press, 2004. 47

[10] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003. 5, 61, 73, 113, 115, 116, 119, 182

[11] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006. 30, 31, 32, 41, 42, 61, 113, 116, 156, 176, 180, 181, 182

[12] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. *When Is "Nearest Neighbor" Meaningful?*, pages 217–235. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. 139, 140

[13] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 217–235, London, UK, UK, 1999. Springer-Verlag. 28, 171

[14] S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 19–26, Washington, DC, USA, 2004. IEEE Computer Society. 47

[15] G. Bisson and C. Grimal. Co-clustering of multi-view datasets: A parallelizable approach. pages 828–833. IEEE, 2012. 136

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 104

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. 91, 94, 104

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 76, 79, 99, 128, 130, 189

[19] M. Brbić and I. Kopriva. Multi-view low-rank sparse subspace clustering. 08 2017. 4, 50

[20] D. Cai and X. He. Manifold adaptive experimental design for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):707–719, 2012. 136

[21] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1548–1560, 2011. 5, 7, 28, 29, 31, 32, 42, 61, 63, 66, 71, 73, 113, 115, 117, 123, 136, 156, 159, 161, 169, 176, 181, 194

[22] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–594, June 2015. 51

[23] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–594, June 2015. 175, 177

[24] M. Á. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the International Conference on Machine Learning*, pages 167–174, 2010. 64

[25] A. Cichocki and A. huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations, 2008. 155, 165, 168, 169

[26] A. Cichocki and A.-H. Phan. Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, 92:708–721, 2009. 163

[27] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994. 30

[28] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, New York, NY, USA, 2001. ACM. 4, 40, 47, 96, 155, 168, 169

[29] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 126–135, 2006. 8, 29, 32, 35, 39, 63, 65, 66, 69, 75, 82, 96, 98, 116, 118, 133, 162, 187, 189, 206

[30] C. H. Q. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):45–55, 2010. 77, 78, 128, 129, 155, 189

[31] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000. 60

[32] C. Fan. Spectral graph theory. *CBMS Regional Conference Series*, 92, 1997. 47

[33] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 41–50, New York, NY, USA, 2005. ACM. 47

[34] H. Gao, F. Nie, and H. Huan. Local centroids structured non-negative matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1905–1911, 2017. 63

[35] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4238–4246, 2015. 4, 153

[36] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 50, 51

[37] B. Geng, D. Tao, C. Xu, L. Yang, and X. Hua. Ensemble manifold regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1227–1233, June 2012. 181

[38] B. Geng, C. Xu, D. Tao, L. Yang, and X. S. Hua. Ensemble manifold regularization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2396–2402, June 2009. xv, 38, 44, 45

[39] T. George and S. Merugu. A scalable collaborative filtering framework based on co-clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 625–628, 2005. 43

[40] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002. 91, 94, 104

[41] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 177

[42] Q. Gu, C. Ding, and J. Han. On trivial solution and scale transfer problems in graph regularized nmf. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1288–1293, 2011. 63, 64, 65, 69, 70, 115, 116

[43] Q. Gu and J. Zhou. Co-clustering on manifolds. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 359–368, 2009. 5, 40, 63, 75, 82, 98, 113, 116, 119, 121, 126, 132, 133, 136, 194

[44] X. He and P. Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003. 71

[45] P. Henderson. Sammon mapping. *Pattern Recognition Letters*, 18(11-13):1307–1316, 1997. 209

[46] D. Hidru and A. Goldenberg. Equinmf: Graph regularized multiview nonnegative matrix factorization. *CoRR*, abs/1409.4018, 2014. 154

[47] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. 154

[48] J. Hou and R. Nayak. Robust clustering of multi-type relational data via a heterogeneous manifold ensemble. In *Proceedings of the International Conference on Data Engineering*, 2015. 4, 5, 8, 28, 35, 43, 52, 61, 64, 66, 75, 81, 82, 113, 114, 115, 116, 117, 118, 121, 123, 131, 133, 134, 136, 159, 161, 162, 169, 183, 191, 194

[49] C.-J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *ACM SIGKDD*, pages 1064–1072, 2011. 155, 156, 163

[50] X. Hu and H. Liu. Text analytics in social media. In *Mining text data*, pages 385–414. Springer, 2012. 92, 101, 102

[51] J. Huang, F. Nie, H. Huang, and C. Ding. Robust manifold nonnegative matrix factorization. *ACM Trans. Knowl. Discov. Data*, 8(3):11:1–11:21, June 2014. 116

[52] Y. Huang and X. Gao. Clustering on heterogeneous networks. *Wiley Int. Rev. Data Min. and Knowl. Disc.*, 4(3):213–233, May 2014. 115

[53] S. F. Hussain, M. Mushtaq, and Z. Halim. Multi-view document clustering via ensemble method. *Journal of Intelligent Information Systems*, 43(1):81–99, 2014. 136

[54] R. Iyer, J. Wong, W. Tavanapong, and D. A. Peterson. Identifying policy agenda sub-topics in political tweets based on community detection. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 698–705. ACM, 2017. 92, 94

[55] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010. 94, 104

[56] G. Jing, H. Jiawei, L. Jialu, and W. Chi. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, pages 252–260. SIAM, 2013. 4, 6, 33, 154, 157, 158, 169, 175, 191

[57] L. Jing, J. Yun, J. Yu, and J. Z. Huang. High-order co-clustering text data on semantics-based representation model. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 171–182, 2011. 61, 81, 96, 136

[58] W. Jing, W. Xiao, T. Feng, L. C. Hong, Y. Hongchuan, and L. Yanbei. *Adaptive Multi-view Semi-supervised Nonnegative Matrix Factorization*, pages 435–444. Springer International Publishing, 2016. 6, 154

[59] S. Kanaan-Izquierdo, A. Ziyatdinov, and A. Perera-Lluna. Multiview and multifeature spectral clustering using common eigenvectors. *Pattern Recognition Letters*, 102:30 – 36, 2018. 4, 49

[60] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, 1990. 29

[61] D. Kuang, C. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012. 97

[62] A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1413–1421. Curran Associates, Inc., 2011. 4, 49

[63] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999. 5, 28, 61, 63, 78, 116, 129, 179

[64] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001. 5, 28, 34, 61, 63, 83, 98, 104, 132, 135, 154, 155, 162, 176, 179

[65] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004. 82, 136

[66] P. Li, J. Bu, C. Chen, and Z. He. Relational co-clustering via manifold ensemble learning. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 1687–1691, 2012. 35, 38, 42, 44, 52, 64, 75, 98, 113, 114, 115, 122, 126, 156, 183

[67] P. Li, J. Bu, C. Chen, Z. He, and D. Cai. Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *IEEE Transactions on Cybernetics*, 43(6):1871 – 1881, December 2013. 116

[68] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2750–2756. AAAI Press, 2015. 49

[69] Z. Li, J. Liu, and H. Lu. Structure preserving non-negative matrix factorization for dimensionality reduction. *Computer Vision and Image Understanding*, 117(9):1175 – 1189, 2013. 64

[70] B. Long, Z. Zhang, X. Wu, and P. S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the International Conference on Machine Learning*, pages 585–592, 2006. 3, 4, 25, 35, 47, 50, 62, 65, 66, 92, 95, 96, 111, 112, 116, 118

[71] K. Luong, T. Balasubramaniam, and R. Nayak. A novel technique of using coupled matrix and greedy coordinate descent for multi-view data representation. In *Proceedings of the 19th International Conference on Web Information Systems Engineering*, 2018. 175, 176, 179, 192

[72] K. Luong and R. Nayak. Learning association relationship and accurate geometric structure for multi-type relational data. In *Proceedings of the International Conference on Data Engineering*, 2018. 93, 95, 97, 98, 99, 113, 116, 118, 121, 161, 183, 191

[73] K. Luong and R. Nayak. *Clustering Multi-View Data Using Non-negative Matrix Factorization and Manifold Learning for Effective Understanding: A Survey Paper*, pages 201–227. Springer International Publishing, 2019. 95, 174, 175, 183

[74] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec. 2007. 116, 194

[75] W. A. Mohotti and R. Nayak. Corpus-based augmented media posts with density-based clustering for community detection. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 379–386. IEEE, 2018. 94

[76] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010. 93, 94, 103

[77] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 849–856, Cambridge, MA, USA, 2001. MIT Press. 47, 48

[78] A. Park, M. Conway, and A. T. Chen. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Computers in human behavior*, 78:98–112, 2018. 92, 94

[79] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explor. Newsl.*, 6(1):90–105, June 2004. 28, 51

[80] Y. Pei, N. Chakraborty, and K. Sycara. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 2083–2089. AAAI Press, 2015. 93, 95

[81] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng. Robust graph regularized nonnegative matrix factorization for clustering. *ACM Trans. Knowl. Discov. Data*, 11(3):33:1–33:30, 2017. 63

[82] L. Qi, Y. Shi, H. Wang, W. Yang, and Y. Gao. Multi-view subspace clustering via a global low-rank affinity matrix. In H. Yin, Y. Gao, B. Li, D. Zhang, M. Yang, Y. Li, F. Klawonn, and A. J. Tallón-Ballesteros, editors, *Intelligent*

Data Engineering and Automated Learning – IDEAL 2016, pages 321–331, Cham, 2016. Springer International Publishing. 28

[83] B. Qian, X. Shen, Y. Gu, Z. Tang, and Y. Ding. Double constrained nmf for partial multi-view clustering. In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7, 2016. 33, 37

[84] M. Qin, D. Jin, K. Lei, B. Gabrys, and K. Musial-Gabrys. Adaptive community detection incorporating topology and content in social networks. *Knowledge-Based Systems*, 161:342 – 356, 2018. 93, 95

[85] N. Rai, S. Negi, S. Chaudhury, and O. Deshmukh. Partial multi-view clustering using graph regularized nmf. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2192–2197, Dec 2016. 33, 34, 176

[86] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. 91, 94, 104

[87] Y. Ruan, D. Fuhry, and S. Parthasarathy. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1089–1098. ACM, 2013. 92

[88] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam. Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 331–340. ACM, 2012. 92

[89] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 102

[90] A. P. L. Samir Kanaan-Izquierdo, Andrey Ziyatdinov. Multiview and multifeature spectral clustering using common eigenvectors. *Pattern Recognition Letters*, 102:31–36, May 2018. 4, 175

[91] M. Schmidt. Least squares optimization with l1-norm regularization, 2005. 70

[92] A. Serra, P. Galdi, and R. Tagliaferri. Multiview learning in biomedical applications. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, pages 265 – 280. Academic Press, 2019. 174

[93] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. 63, 64, 67, 69, 206

[94] S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23:2031–2038s, 12 2013. 174

[95] L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data mining and knowledge discovery*, 25(1):1–33, 2012. 93, 94, 103

[96] J. B. Tenenbaum. Mapping a manifold of perceptual observations. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 682–688. MIT Press, 1998. 30, 180

[97] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 30, 41, 44, 61, 115

[98] N. T. Trendafilov. Stepwise estimation of common principal components. *Comput. Stat. Data Anal.*, 54(12):3446–3457, Dec. 2010. 50

[99] H. Wang, H. Huang, and C. Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 279–284, 2011. 3, 4, 28, 35, 37, 41, 44, 61, 62, 64, 66, 75, 82, 92, 95, 96, 98, 103, 112, 113, 114, 115, 116, 117, 118, 121, 122, 123, 126, 133, 136

[100] H. Wang, F. Nie, H. Huang, and C. Ding. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In *Proceedings the*

International Conference on Data Mining, pages 774–783, 2011. 37, 69, 113, 114, 115, 122

[101] H. Wang, Y. Yang, and T. Li. Multi-view clustering via concept factorization with local manifold regularization. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1245–1250, Dec 2016. xv, 3

[102] J. Wang, F. Tian, X. Wang, H. Yu, C. H. Liu, and L. Yang. Multi-component nonnegative matrix factorization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2922–2928, 2017. 6, 34, 154, 157, 158, 179

[103] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang. Diverse non-negative matrix factorization for multiview data representation. *IEEE Transactions on Cybernetics*, PP(99):1–13, 2017. 4, 6, 34, 153, 154, 158, 175, 177, 179, 191, 192

[104] L. Wang, D. Li, T. He, and Z. Xue. Manifold regularized multi-view subspace clustering for image representation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 283–288, Dec 2016. 4

[105] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, Nov 2015. 4, 51

[106] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, Nov 2015. 113

[107] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013. 4, 26, 174, 175

[108] J. Xu, J. Han, and F. Nie. Multi-view feature learning with discriminative regularization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3161–3167, 2017. 5, 7

[109] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 267–273, 2003. 29, 63

[110] K. Zhan, J. Shi, J. Wang, and F. Tian. Graph-regularized concept factorization for multi-view document clustering. 48:411–418, 2017. 5, 7

[111] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Transactions on Image Processing*, 16(1):57–68, Jan 2007. 190

[112] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2):648–659, Feb 2017. 177

[113] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao. Latent multi-view subspace clustering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4333–4341, July 2017. 50

[114] X. Zhang, H. Gao, G. Li, J. Zhao, J. Huo, J. Yin, Y. Liu, and L. Zheng. Multi-view clustering based on graph-regularized nonnegative matrix factorization for object recognition. 11 2017. 33, 34, 39

[115] X. Zhang, H. Gao, G. Li, J. Zhao, J. Huo, J. Yin, Y. Liu, and L. Zheng. Multi-view clustering based on graph-regularized nonnegative matrix factorization for object recognition. *Information Sciences*, 432:463 – 478, 2018. 116, 176, 179, 180, 183

[116] X. Zhang, H. Li, W. Liang, and J. Luo. Multi-type co-clustering of general heterogeneous information networks via nonnegative matrix tri-factorization. pages 1353–1358, 2016. 61, 153

[117] X. Zhang, L. Zhao, L. Zong, X. Liu, and H. Yu. Multi-view clustering via multi-manifold regularized nonnegative matrix factorization. In *Proceedings*

of the International Conference on Data Mining, pages 1103–1108, 2014. 4, 28, 37, 38, 45, 46, 114, 115, 133, 135

[118] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview. *Inf. Fusion*, 38(C):43–54, Nov. 2017. 154

[119] S. Zhong and J. Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384, 2005. 82, 100, 136, 169, 190

[120] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao. Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, 88(Supplement C):74 – 89, 2017. 3, 4, 52, 112, 113, 116, 118, 153, 154, 156, 158, 176, 177, 179, 180, 184, 191

[121] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao. Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, 88:74 – 89, 2017. 33, 64