

Blekinge Institute of Technology  
Doctoral Dissertation Series No. 2024:XX  
ISSN 1653-2090  
ISBN xxx-xx-xx-x

# Mining Evolving and Heterogeneous Data

Cluster-based Analysis Techniques

**Vishnu Manasa Devagiri**



DOCTORAL DISSERTATION  
for the degree of Doctor of Philosophy at Blekinge Institute of Technology to be publicly  
defended on [XXXXXXX] at [time] in [xx Hall, Address]

Supervisors  
Xxxxxxx xxxxxxxx, XXXXXXXX, XXXXXXXXXX  
Xxxxxxx xxxxxxxx, XXXXXXXX, XXXXXXXXXX

Faculty Opponent  
Xxxxxxx, xxxxxxxx, XXXXXXXX, XXXXXXXXXX

# Abstract

A large amount of data is generated from fields like IoT, smart monitoring applications, etc., raising demand for suitable data analysis and mining techniques. Data produced through such systems have many distinct characteristics, like continuous generation, evolving nature, multi-source origin, and heterogeneity, which are usually unannotated. Clustering is an unsupervised learning technique used to group and analyze unlabeled data. Conventional clustering algorithms are unsuitable for dealing with data with the mentioned characteristics due to memory, computational constraints, and their inability to handle heterogeneous and evolving nature. Therefore, novel clustering approaches are needed to analyze and interpret such challenging data.

This thesis focuses on building and studying advanced clustering algorithms that can address the main challenges of today's real-world data: evolving and heterogeneous nature. An evolutionary clustering approach capable of continuously updating the generated clustering solution in the presence of new data is initially proposed, which is later extended to address the challenges of multi-view data applications. Multi-view or multi-source data presents the studied phenomenon or system from different perspectives (views) and can reveal interesting knowledge that is not visible when only one view is considered and analyzed. This has motivated us to continue exploring data from different perspectives in several other studies of this thesis. Domain shift is a common problem when data is obtained from various devices or locations, leading to a drop in the performance of machine learning models if they are not adapted to the current domain (device, location, etc.). The thesis also explores the domain adaptation problem in a resource-constraint way using the cluster integration techniques proposed. A new hybrid clustering technique for analyzing heterogeneous data, which produces homogeneous groups facilitating continuous monitoring and fault detection, is also proposed.

The algorithms or techniques proposed in this thesis are evaluated on various data sets, including real-world data from industrial partners in domains like smart building systems, smart logistics, and performance monitoring of industrial assets. The obtained results demonstrated the robustness of the algorithms for modeling, analyzing, and mining evolving data streams and/or heterogeneous data. They can adequately adapt single and multi-view clustering models by continuously integrating newly arriving data.

**Keywords:** Multi-view clustering, Evolutionary clustering, Streaming data, Heterogeneous data

Blekinge Institute of Technology  
Doctoral Dissertation Series No. 2024:XX

# Mining Evolving and Heterogeneous Data

Cluster-based Analysis Techniques

**Vishnu Manasa Devagiri**

Doctoral Dissertation in Computer Science



Department of Computer Science  
Blekinge Institute of Technology  
SWEDEN

Copyright pp Vishnu Manasa Devagiri  
Paper 1 © ...  
Paper 2 © ....  
Paper 3 © ...  
Paper 4 © by the Authors (Manuscript unpublished)


Blekinge Institute of Technology  
Department of Computer Science

Blekinge Institute of Technology Doctoral Dissertation Series No. 2024:XX  
ISBN xxx-xx-xx-x  
ISSN 1653-2090  
urn:nbn:se:bth-????

Printed in Sweden by Media-Tryck, Lund University, Lund 2024



Media-Tryck is a Nordic Swan Ecolabel  
certified provider of printed material.  
Read more about our environmental  
work at [www.mediatryck.lu.se](http://www.mediatryck.lu.se)

**MADE IN SWEDEN** 

*Dedication*

*This is a dedication.*



# Acknowledgements

This is an acknowledgement





# List of Papers

## Paper I

V. Boeva, M. Angelova, V. M. Devagiri, and E. Tsiporkova. “Bipartite Split-Merge Evolutionary Clustering”. In: *Agents and Artificial Intelligence*. Ed. by J. van den Herik, A. P. Rocha, and L. Steels. Cham: Springer International Publishing, 2019, pp. 204–223. DOI: 10.1007/978-3-030-37494-5\_11

## Paper II

V. M. Devagiri, V. Boeva, and E. Tsiporkova. “Split-Merge Evolutionary Clustering for Multi-View Streaming Data”. In: *Procedia Computer Science* 176 (2020). Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020, pp. 460–469. ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.08.048

## Paper III

V.M. Devagiri, V. Boeva, and S. Abghari. “A Multi-view Clustering Approach for Analysis of Streaming Data”. In: *Artificial Intelligence Applications and Innovations*. Ed. by I. Maglogiannis, J. Macintyre, and L. Iliadis. Cham: Springer International Publishing, 2021, pp. 169–183. ISBN: 978-3-030-79150-6. DOI: 10.1007/978-3-030-79150-6\_14

## Paper IV

V. M. Devagiri, V. Boeva, S. Abghari, F. Basiri, and N. Lavesson. “Multi-View Data Analysis Techniques for Monitoring Smart Building Systems”. In: *Sensors* 21.20 (2021). ISSN: 1424-8220. DOI: 10.3390/s21206775

## Paper V

C. Åleskog, V. M. Devagiri, and V. Boeva. “A Graph-Based Multi-view Clustering Approach for Continuous Pattern Mining”. In: *Recent Advancements in Multi-View Data Analytics*. Ed. by W. Pedrycz and S.-M. Chen. Cham: Springer International Publishing, 2022, pp. 201–237. ISBN: 978-3-030-95239-6. DOI: 10.1007/978-3-030-95239-6\_8

## Paper VI

V. M. Devagiri, V. Boeva, and S. Abghari. “Domain Adaptation Through Cluster Integration and Correlation”. In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2022, pp. 1–8. DOI: 10.1109/ICDMW58026.2022.00025

## Paper VII

V. M. Devagiri, V. Boeva, and S. Abghari. ”A Domain Adaptation Technique through Cluster Boundary Integration”. *Evolving Systems*. (minor revision was submitted on 04 January 2024).

## Paper VIII

V. M. Devagiri, P. Dagnely, V. Boeva, and E. Tsiorkova. ”Putting Sense into Multi-source Heterogeneous Data with Hypergraph Clustering Analysis”. Accepted for Symposium on Intelligent Data Analysis (IDA), Stockholm, Sweden, April 2024.

Other publications related but not included in the thesis are:

## Paper I

M. Angelova, V. M. Devagiri, V. Boeva, P. Linde and N. Lavesson. ”An Expertise Recommender System based on Data from Institutional Repository (DiVA)”. Leslie Chan and Pierre Mounier (Eds.): *Connecting the Knowledge Commons – from projects to sustainable infrastructure*. OpenEdition Press, pp.135-149, 2019. DOI: 10.4000/books.oep.9078

## Paper II

V. Boeva, M. Angelova, V. M. Devagiri, E. Tsiporkova, A Split-Merge Framework for Evolutionary Clustering, 31th Swedish AI Society Workshop SAIS 2019, Umeå, Sweden, June 2019.

## Paper III

V. Boeva, E. Casalicchio, S. Abghari, A.A. Al-Saedi, V. M. Devagiri, A. Petef, P. Exner, A. Isberg. and M. Jasarevic. 2022. "Distributed and Adaptive Edge-based AI Models for Sensor Networks (DAISeN)". Position Papers of the 17th Conference on Computer Science and Intelligence Systems, Annals of Computer Science and Information Systems 31 (2022): 71-78. DOI: 10.15439/2022F267

## Funding

The research work done as a part of this thesis is partially funded by the following:

- "Scalable resource efficient systems for big data analytics", project funded by the Swedish Knowledge Foundation (grant: 20140032).
- "Distributed and Adaptive Edge-based AI Models for Sensor Networks", Sony Research Award Program 2020 Project.
- "Human-centered Intelligent Realities (HINTS)", project funded by the Swedish Knowledge Foundation (grant: 20220068).

## Author's contribution to the papers

The author is the main driver and the first author for all the papers except for papers I and V. For the studies where she was the main driver and first author, she was involved in all the phases of the research, that is, idea generation, designing and conducting experimentation, analysis of results, writing the original draft, reviewing and editing the manuscript. For Paper I, she was mainly involved in designing and conducting experiments, analyzing results, reviewing and editing the manuscript. For Paper V, the author was involved in the idea generation, experimental design, analyzing results, reviewing and editing the manuscript. The author was also the co-supervisor for this study. The formatting of the papers is changed to be adapted to the thesis.

# Abbreviations

FCA	Formal Concept Analysis.
HAR	human activity recognition.
ML	Machine Learning.



# Table of Contents

<b>Acknowledgements</b>	<b>i</b>
<b>List of Papers</b>	<b>iii</b>
<b>Abbreviations</b>	<b>vii</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Research Problem . . . . .	2
1.2 Contributions and Papers Included . . . . .	4
1.3 Thesis Structure . . . . .	6
<b>Chapter 2 Background and Related Work</b>	<b>9</b>
2.1 Evolutionary Clustering . . . . .	9
2.1.1 Concept Drift . . . . .	9
2.2 Stream Clustering . . . . .	10
2.3 Multi-View Clustering . . . . .	10
2.4 Multi-View Stream Clustering . . . . .	11
2.5 Graph-based Clustering . . . . .	11
2.6 Multi-Instance Clustering . . . . .	11
2.7 Formal Concept Analysis . . . . .	12
<b>Chapter 3 Methodology</b>	<b>13</b>
3.1 Research Questions . . . . .	13
3.2 Data sets . . . . .	15
3.3 Research Methodology . . . . .	15
3.4 Evaluation Metrics . . . . .	17
3.5 Validity Threats . . . . .	17
3.5.1 Internal Validity Threat . . . . .	18
3.5.2 External Validity Threat . . . . .	18
3.5.3 Construct Validity Threat . . . . .	18
<b>Chapter 4 Results and Analysis</b>	<b>19</b>
4.1 Research Questions' Answers . . . . .	20
<b>Chapter 5 Conclusions and Future Work</b>	<b>23</b>
<b>Chapter 6 Experiences and Learning Outcomes</b>	<b>25</b>
<b>Bibliography</b>	<b>27</b>





# 1 Introduction

A lot of data is available today, thanks to the growth and recent advancements in fields like the Internet of Things (IoT), sensor networks, smart monitoring applications, etc. These applications generate data continuously, and there are various challenges in storing, processing, and obtaining valuable information from such huge amounts of data [1]. Machine Learning (ML) and data mining fields provide methods and techniques that can be applied to analyze data for extracting useful knowledge and insights that can be used to understand or monitor the studied system. There are various ML and data mining algorithms that can be broadly categorized as supervised, semi-supervised, and unsupervised. In the current research environment, supervised techniques are much more evolved when compared to the other two areas. Both supervised and semi-supervised learning algorithms require a large amount of labeled data for training, which is either generally unavailable when dealing with real-world data and/or is a costly process to label such large volumes of data. Unsupervised learning techniques have a greater demand and use in real-world scenarios, as they do not require labeled data. Clustering is one of the most popular unsupervised learning techniques [2]. Clustering techniques are used to group data such that data points placed in a cluster are similar to each other and different from the data points of other clusters [3]. This thesis is specially focused on clustering techniques that handle evolving and/or multi-source/view data.

Data is generated continuously in many application scenarios, and the machine learning models built become obsolete over time due to changes in data characteristics. This provides the need for evolutionary algorithms, which can be updated at regular intervals to be suitable for new data. Many computational resources are required to rebuild the model every time new information is generated. Hence, evolutionary clustering algorithms are required to discover and accommodate data with new characteristics. In addition to the streaming nature and concept drift, there is also a need to develop clustering algorithms that can handle data generated from multiple sources, as most applications, like smart monitoring systems, collect information from various devices. Such data generated from multiple sources (also known as multi-view data) observing the same event can present interesting details that are otherwise not visible when studying data from a single source [4]. Data heterogeneity is an additional problem that needs to be addressed when working with multi-view data as it is collected from different sources, e.g., multiple sensors or devices [5].

Traditional clustering algorithms are unsuitable for addressing the above-stated challenges of multi-source [6] and streaming [7] data. In this thesis, novel clustering approaches are designed to address different aspects of the above-discussed challenges. The developed algorithms can monitor, analyze, and interpret the data obtained from different smart applications used for monitoring, providing personalized recommendations, etc.

## 1.1 Research Problem

This thesis combines several research works presenting robust clustering techniques suitable for analyzing and extracting knowledge from evolving and heterogeneous data from single and multiple sources. Eight different studies have been conducted in this PhD thesis. The main aim of the thesis is to develop clustering algorithms suitable to mine and analyze evolving and heterogeneous streaming data generated from single or multiple sources. The following research objectives are set to achieve the thesis goal.

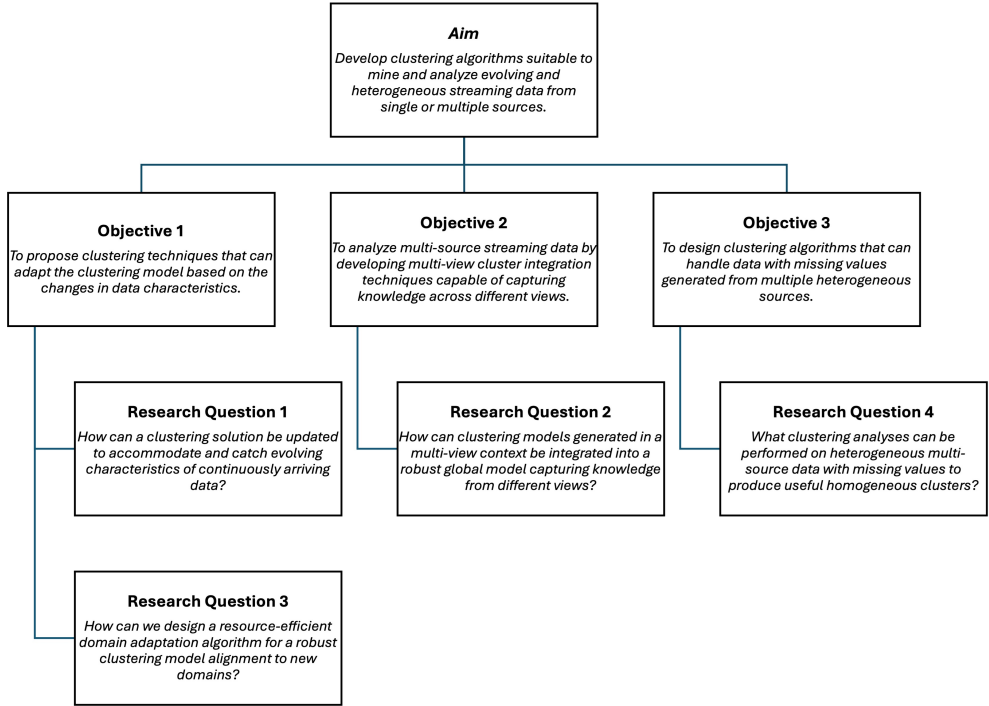
- Obj 1. To propose clustering techniques that can adapt the clustering model based on the changes in data characteristics.
- Obj 2. To analyze multi-source streaming data by developing multi-view cluster integration techniques capable of capturing knowledge across different views.
- Obj 3. To design clustering algorithms that can handle data with missing values generated from multiple heterogeneous sources.

Let by the aims and objectives set to be accomplished, the following research questions are framed, addressed, and answered in this thesis. A visualization of the connections between the aim, objectives, and research questions is presented in Figure 1.1.

**RQ1** *How can a clustering solution be updated to accommodate and catch evolving characteristics of continuously arriving data?*

**Motivation:** In many of the current-day applications, lots of data is continuously generated whose characteristics tend to change over time. In this context, the new data does not correctly fit into the current model, and it becomes difficult to accommodate it. At the same time, we cannot rebuild the clustering model from scratch each time new data arrives, as this would consume lots of resources. Hence, there is a need for a clustering approach that can successfully integrate the new data into the existing model by considering its new characteristics.

**RQ2** *How can clustering models generated in a multi-view context be integrated into a robust global model capturing knowledge from different views?*



**Figure 1.1:** Figure visualizing the connections between the aim, objectives, and research questions of the thesis.

**Motivation:** Many applications these days generate multi-source data, and information from each of these sources is complementary. When viewed together, this data can give or produce information that cannot be obtained when only each of these views is considered [4]. Therefore, successfully integrating this knowledge from multiple views into a single model might be helpful. Even though there have been lots of works in multi-view clustering and stream clustering, not much has been done in the area of multi-view stream clustering [8, 9].

**RQ3** *How can we design a resource-efficient domain adaptation algorithm for a robust clustering model alignment to new domains?*

**Motivation:** There have been a lot of works in the field of domain adaptation, but the majority of the state-of-the-art works are done through deep learning or are developed for computer vision-based applications. There are limited works addressing this aspect for time-series data. Such data is usually generated by IoT devices, which have strict resource constraints. This raises the demand for resource-efficient algorithms.

**RQ4** *What clustering analyses can be performed on heterogeneous multi-source data*

*with missing values to produce useful homogeneous clusters?*

**Motivation:** When data is generated from multiple sources, it is likely that the data is heterogeneous. In such data, certain interesting aspects might only be associated with a certain type of features, and traditional clustering algorithms are not very suitable for grouping the data. Using multi-view techniques to analyze a few features at a time might be helpful. Additionally, with heterogeneous data, it becomes difficult to analyze and derive meaningful insights or perform system monitoring; producing homogeneous groups with similar properties might make this task easier.

## 1.2 Contributions and Papers Included

The main contribution of this thesis is the development of novel clustering techniques suitable for the mining and analysis of evolving and heterogeneous (multi-source/view) data. This thesis includes eight papers, five of which, namely Papers II, III, IV, V, and VIII, deal with heterogeneity, and all the others except Paper VIII are designed to address challenges related to the evolving nature and domain shift (when multiple devices or locations are considered) of the streaming data. An evolutionary clustering approach, entitled Split-Merge Clustering, capable of continuously updating the generated clustering solution in the presence of new data, is proposed in Paper I. New challenges have been studied and addressed through the progression of the work. Namely, the Split-Merge Clustering algorithm has been enhanced in Paper II to deal with the challenges of multi-view data applications. Multi-view or multi-source data presents the studied phenomenon/system from different perspectives (views) and can reveal interesting knowledge that is not visible when only one view is considered and analyzed. This has motivated us to continue exploring this in a few other studies (III, IV, V). The algorithm proposed in Paper III improves the performance and interpretability of the algorithm proposed in Paper II. Paper IV introduces a minimum spanning tree-based multi-view clustering algorithm capable of transferring knowledge between consecutive data chunks, and it is also enriched with a post-clustering pattern-labeling procedure. In Paper V, the algorithm proposed in Paper III is evaluated in an industrial use case of smart buildings, and a context-aware approach and extension to integrated system analysis are presented. Papers VI and VII propose a clustering technique to deal with the domain adaptation problem. Paper VIII presents a novel hybrid clustering technique for grouping heterogeneous data to obtain homogeneous groups suitable for continuous monitoring and deviating behavior detection.

The following text briefly summarizes the papers included in this thesis, along with the contributions of each of these.

*Paper I.* A novel clustering approach entitled *Split-Merge Evolutionary clustering* is proposed. The approach integrates the clustering models of historical

and newly arriving data using a bipartite graph obtained based on the correlations between the two clustering models. An updated clustering solution is obtained by splitting or merging the clusters based on the edge connections of the bipartite graph. The algorithm is evaluated on four different data sets and compared with two other state-of-the-art algorithms.

*Paper II.* A multi-view clustering approach, *MV Split-Merge Clustering* based on the algorithm in Paper I is proposed. It is designed to analyze multi-view streaming data and build a consensus clustering solution (global model) based on the information obtained from different views where Formal Concept Analysis (FCA) is used for the integration. An initial evaluation is done, the algorithm is compared to its batch version and it has produced comparable results on an anthropometric data set, showcasing the algorithm's potential as a multi-view clustering solution.

*Paper III.* A novel multi-view clustering algorithm, entitled *MV Multi-Instance clustering* is proposed. The new algorithm is designed to provide improved performance and interpretability of results when compared to *MV Split-Merge Clustering*. Unlike Paper II, this paper uses a novel multi-instance learning algorithm proposed, *Bi-Correlation MI-Clustering* instead of *MV Split-Merge clustering* to update clustering solutions in each view. Apart from this, closed patterns are used to mine frequent concepts while building the global model, which reduces its complexity. The results show that the proposed algorithm has performed better than the *MV Split-Merge clustering*.

*Paper IV.* This work studies the use of *MV Multi-Instance Clustering* algorithm for multi-view analysis of data in the smart building domain, using data provided by one of our industrial partners. The scenarios in which the algorithm could be used to analyze the data are presented and examined, focusing on contextual and integrated analysis of the systems. It also presents visualization techniques to showcase extracted knowledge that could be used to aid domain experts in detecting trends. The study showed the algorithm's potential in monitoring, analyzing, and identifying deviating behaviors of sub-systems in a smart building system.

*Paper V.* A novel multi-view clustering algorithm entitled *MST-MVS* is proposed. It is based on minimum spanning tree (MST) clustering and can be used to analyze and monitor streaming data. It is a continuous data mining approach where the integrated knowledge from the global model obtained at each data chunk is transferred to the next one through artificial nodes of the MST algorithm. This knowledge transfer which proved to be beneficial, helps to build the clustering solution in the next chunk by taking into account how the previous data has been grouped. A post-labeling technique used to label the chunk's data points based on the built global

model is also proposed. *MST-MVS* is evaluated on both real-world and synthetic data sets.

*Paper VI.* This work proposes a resource-efficient novel domain adaptation technique based on cluster correlation and integration, entitled *DIBCA*. It integrates knowledge from different domains, i.e., the source and target domain, by identifying their correlations. Correlations are obtained by labeling source data with the target model and target data with the source model. *DIBCA* produces an integrated model that can be used across the domains and a personalized adapted model for each domain. Its capability in automatic data labeling is showcased using the human activity recognition (HAR) data set, and in addition, a real-world industrial use case of smart logistics is used to study its potential in the domain adaptation task.

*Paper VII.* This study proposes *DIBCA++*, a novel domain adaptation technique and an improved version of *DIBCA*. *DIBCA++* is designed to be robust to outliers compared to its predecessor. It requires a modest amount of storage and computational resources as it uses the clusters' mean, standard deviation, and size. The explainability aspects and applicability potential of the algorithm are also studied and presented. The algorithm is evaluated on a HAR data set and a smart logistics use case from our industrial partner. The experimental results showcase the better performance of *DIBCA++* over *DIBCA*. They also present the ability of *DIBCA++* to transfer knowledge between domains.

*Paper VIII.* In this study, a novel unsupervised data analysis method is proposed to group heterogeneous data with missing values into homogeneous groups that can be used for performance monitoring. Each group is expected to have comparable behavior, thus aiding domain experts in monitoring the performance of assets in the group. The proposed approach is designed such that it can handle missing values. The approach is based on concepts of multi-layer clustering, shared nearest-neighbor similarity, and hyper-graph clustering. The proposed approach is evaluated on a real-world industrial data set of multi-source heterogeneous assets (compressors) with a substantial amount of missing values.

### 1.3 Thesis Structure

The rest of the thesis is structured as follows. Chapter ?? presents the background required for an easy understanding of the concepts used in different studies of the thesis along with a summary of the works related to the thesis. This is followed by Chapter ??, which offers an overview of the scientific methodology used in this

thesis. It covers research questions, data sets used, evaluation metrics, and validity threats. The results of the thesis and their analysis are presented in Chapter ?? and the conclusions and future work are presented in Chapter 5. In Chapter ??, experiences and the learning outcomes identified during the PhD journey are presented. The next eight Chapters consist of the research papers included in this thesis.





# 2 Background and Related Work

A brief introduction to different concepts related to the thesis and works done in these areas is presented in this chapter.

## 2.1 Evolutionary Clustering

Evolutionary clustering algorithms are designed to continuously accommodate data with evolving characteristics, i.e., the data characteristics might change over time. Section 2.1.1 presents more information about this phenomenon. Clustering is an unsupervised learning technique where labeled data is not required to build a model. Similar to the traditional clustering algorithms, the main aim of these clustering algorithms is to group the data, where data points belonging to a group are similar to each other and are different from the data points of other groups. According to Bouchachia [10], different phases of an evolutionary clustering algorithm can be categorised into matching, accommodating new data, and model refinement.

In papers [11–13] the authors propose novel clustering algorithms capable of splitting or merging the clusters based on the need. Lughofer proposes a dynamic clustering algorithm entitled *Dynamic split-merge* algorithm [11]. The algorithm is designed to be used in integration with another incremental clustering algorithm. Incoming data points are initially accommodated into the existing clustering solution; this is followed by evaluating each cluster to determine if it needs to be split, merged, or retained.

### 2.1.1 Concept Drift

Concept drift is a phenomenon where the data characteristics tend to change over time [7]. Thereby making the ML model outdated and degrading its performance if concept drift is not handled [14]. Such changes in data characteristics are common in many fields where data is generated over time. For example, if we consider consumer profiles of a clothing store, what a person buys from that store might change over a period of time due to some external factors like climate, buying gifts, etc. Such

changes in data characteristics is known as concept drift. There are different types of concept drift scenarios introduced in the literature, such as blip, gradual, incremental, noise, recurring, or sudden [15].

## 2.2 Stream Clustering

As the name suggests, stream clustering techniques are used to cluster streaming data. Such type of data is continuously generated. Due to the large amounts of data produced, streaming data is generally not labeled. Hence, clustering is one of the most suitable techniques to analyze or mine useful information from such data [7]. While designing stream clustering algorithms, it is essential that the time and memory constraints are considered for handling large amounts of data. Concept drift, defined in Section 2.1.1 is a common problem that needs to be addressed by a clustering algorithm designed for streaming data [7].

In [7, 16, 17], the authors review the existing state-of-the-art stream clustering algorithms. An iterative Non-negative matrix factorization based algorithm is proposed in [18], entitled as ONMF. Another stream clustering algorithm, SVStream is proposed in [19], which is a predecessor of MVStream [8], a multi-view stream clustering algorithm.

## 2.3 Multi-View Clustering

Multi-view clustering techniques are used to cluster data obtained from multiple sources. Viewing the information from numerous sources can sometimes provide valuable information that is not obvious when viewing it from just one source [4]. For example, a patient profile can be represented by the data collected from different smart monitoring devices, notes written by doctors, images such as x-rays, etc. A multi-view clustering technique considers information from across the views and builds a consensus or aggregated model representing this information [20].

Many challenges need to be addressed while working with data obtained from multiple sources. As the data is obtained from various sources, there is possibility that it could be heterogeneous [21]. Incomplete views, that is, the possibility of missing information from different views is another common challenge. Authors of [9, 22, 23] address the later challenge in their works.

Studies [5, 6] are surveys providing a good overview of the recent works in the field. Some of the novel algorithms proposed in the field are briefly presented in the following text. Wang et al. [24] in their work propose a novel algorithm entitled *MVC-LFA*. In the conducted experiments, the proposed algorithm performed better than the other considered algorithms. A multi-view clustering algorithm based on non-negative matrix factorization is proposed by Shao et al. in [9]. Zhu et al. [25]

have proposed another algorithm in the field that is based on feature selection.

## 2.4 Multi-View Stream Clustering

As the name suggests, multi-view stream clustering algorithms are expected to address the challenges of both multi-view and stream clustering algorithms. That is it should be able to analyze and group streaming data generated from more than one source. Compared to the fields of multi-view and stream clustering, multi-view stream clustering is still in its infancy [8, 9]. Authors of [8, 9] have proposed novel algorithms in the field. MVStream proposed in [8] is based on support vectors, and is robust to concept drift. In [9], online multi-view clustering based on non-negative matrix factorization is proposed.

## 2.5 Graph-based Clustering

This section gives a brief over of the graph-based algorithms used in our work. Bipartite graphs and Cut clustering based on Minimum Spanning Tree (MST), which are used as a part of our proposed algorithms, are presented in this section.

As stated in [26], there are many ways to cluster data from a bipartite graph. The authors define Bipartite correlation clustering as follows. They state that a bi-clique of the bipartite graph represents a cluster, and all the bi-cliques together represent the clustering solution. Ailon et al. [26] propose *PivotBiCluster*, a bipartite correlation clustering algorithm which does not need to have prior knowledge about the number of clusters. The algorithm uses merge functionality to merge clusters from either sides to obtain the final clustering solution.

In this thesis *Cut-clustering* algorithm [27], that uses minimum cuts in a graph to cluster data is used. The algorithm considers the data instances as nodes of the graph. An undirected graph with the edge weight representing the similarity between the nodes is built. This is followed by introducing an artificial node connected to all the other nodes of the graph with a constant distance  $\alpha$ . Then, the graph's Minimum Spanning Tree (MST) is computed, followed by removing the artificial node. The forest obtained after removing the artificial node is the final clustering solution obtained. Clustering algorithms proposed in [28, 29] use this *Cut-clustering* algorithm.

## 2.6 Multi-Instance Clustering

In Multi-instance (MI) learning, unlike traditional learning algorithms, a group of instances (also known as a bag) are considered as a single data object [30]. MI clustering is an unsupervised learning technique used to group these bags into clusters.

Bags within the same clusters are similar to each other and are different from the bags in the other clusters. Like traditional clustering algorithms, MI clustering does not require labeled data and can create the groups based on the built-in data structure. Even though the basic properties of MI clustering are similar to traditional clustering algorithms, it cannot be entirely treated like them as a single data instance considered here has many instances that might have different characteristics [31].

## 2.7 Formal Concept Analysis

Formal Concept Analysis (FCA) [32] is a technique used to find relations between objects and their properties, also known as attributes in ML terminology. It has been used in areas like data mining, machine learning to extract useful information. Formal context and concept lattice are two important parts of FCA. Formal context is a table where the rows and columns consist of objects and attributes, respectively. If an object possesses an attribute, the cell corresponding to this object and attribute is marked with a cross. Concept lattice, a hierarchical structure, is derived from the formal context. The concept lattice contains concepts that can be represented using  $(X, Y)$ , where  $X$  and  $Y$  belong to the subset of objects and attributes, respectively. A concept represents a group of data points all sharing the properties (attributes) present in the concept and vice-versa. In the concept hierarchy, there exists a super and sub concept for each concept. Previously many works such as [33–35] used FCA as a part of their clustering approaches to analyze or aggregate the clustering solutions.

# 3 Methodology

This thesis is related to the areas of data mining, knowledge discovery, and ML; it explores unsupervised learning techniques in the areas of evolutionary clustering, multi-view clustering, and domain adaptation. The main aim of the thesis is to develop clustering algorithms suitable for the mining and analysis of evolving and heterogeneous streaming data generated from single or multiple sources.

## 3.1 Research Questions

This section presents the research questions of the thesis along with a brief description of how each question is addressed.

**RQ1** *How can a clustering solution be updated to accommodate and catch evolving characteristics of continuously arriving data?*

This research question is addressed in papers I, II, III, IV, and V. In Papers I and II, our proposed *Split-Merge evolutionary clustering* algorithm has been used to accomplish this task of successfully integrating newly arriving data into the existing clustering solution. The algorithm is based on bipartite correlation clustering. The correlations between the existing clustering model and the clustering solution of the newly arriving data chunk are considered to determine the similarities and differences between the two clustering solutions. Based on the correlations, the existing clusters are either merged, split, or retained as they were. In Paper III, we have proposed an evolving clustering approach entitled *Bi-Correlation MI-Clustering* based on MI clustering and bipartite correlation clustering. In the previous algorithm, only representatives from the clusters were used to find the correlations, which might sometimes impact the performance. The proposed algorithm in Paper III overcomes this by using MI-Clustering; each cluster is treated as an individual data object (bags of instances). In Paper IV, the proposed *Bi-Correlation MI-Clustering* is used in the real-world use case of monitoring smart systems in smart building domains. Paper V proposes yet another evolving clustering approach based on MST clustering. Unlike others, this algorithm uses knowledge from the global model (model integrating knowledge from all the views) when building the clustering solution of the next data chunk.

**RQ2** *How can clustering models generated in a multi-view context be integrated into a robust global model capturing knowledge from different views?*

Papers II, III, IV, and V address this research question. In both Papers II, III, and IV, FCA has been used to integrate the knowledge from different views into a global model consisting of a formal context and concept lattice. In addition to FCA, closed patterns are used in papers III and IV to extract frequently occurring patterns from the formal context, thereby reducing the complexity of the concept lattice. It can be noted that Paper IV uses the algorithm proposed in Paper III and showcases its potential in analyzing and monitoring the sub-systems of the smart building domain. A different integration approach based on transferring the representatives of each cluster from all the views is used in Paper V. For each of the representatives being used, the values of their attributes from other views are also extracted. An integrated matrix is built using the representatives and their extracted attributes' values. This is followed by using MST based clustering algorithm to cluster the data points available in the integrated matrix to obtain the final global model. In Paper V, the local clustering models of each view are evaluated before being used to build the global model. Only the views whose local clustering models have an SI score greater than the predefined threshold are used to build the global model. This evaluation is done to make sure that the quality of the local clustering model does not negatively affect the global model. It is important that the global model is of good quality as the knowledge obtained from this is transferred and used to seed the local clustering model of the next data chunks. A specifically devoted experiment has studied this.

**RQ3** *How can we design a resource-efficient domain adaptation algorithm for a robust clustering model alignment to new domains?*

Papers VI and VII present domain adaptation algorithms suitable for robust cluster model alignment to new domains and address this research question. Both these algorithms adapt clustering solutions from different domains by identifying correlations across models and then performing cluster integration. Cross-labeling of models across domains (the model of one domain is used on the representatives of the other domain) is used to identify the correlations. The algorithms produce an integrated clustering model that can be used across the domains and adapted domain models, one for each domain. Both the algorithms are designed to be resource efficient as all the operations are performed using only the representatives. Paper VI proposes *DIBCA*, which uses the clusters' low-value and high-value vectors as representatives. These representatives are used for all the operations by the algorithm. In Paper VII, *DIBCA++* an optimized version of *DIBCA*, which is more robust to outliers is proposed. *DIBCA++* uses the clusters' mean, standard deviation, and size as representatives.

**RQ4** *What clustering analyses can be performed on heterogeneous multi-source data with missing values to produce useful homogeneous clusters?*

This research question is addressed in paper VIII. A multi-layered clustering approach followed by hypergraph clustering based on k-medoids and nearest-neighbor similarity is used to achieve this. The available feature set is initially divided into different layers,

## 3.2 Data sets

The data sets used for evaluating the algorithms proposed in this thesis are presented in this section. We have used a mix of synthetic data, publicly available real-world data, and also data from real-world industrial use cases from our industrial partners. We have used nine different types of data sets: cover type [36], yeast [37], wine quality [38], anthropometric data [39], real-time sensor data, Dim32 [40], ... and .. across papers. Data sets, cover type, yeast, and wine quality, are obtained from the UCI ML repository. The anthropometric data set is a publicly available data set of undergraduate students. It classifies whether a person has high blood pressure or not based on their anthropometric measures. The real-time sensor data is obtained from a smart building company based in Stockholm, Sweden. Table 3.1 presents the details about the data sets used in different papers like the number of samples used in the experimentation, number of attributes, and classes available.

**Table 3.1:** Data sets used in the thesis.

Data set	Samples	Attributes	Classes	Papers
Cover-type	50000	14	7	I, V
Yeast	1484	8	10	I
Wine quality	6498	12	7	I
anthropometric data	400	9	6	I, II, III
Sensor data - Smart building Domain	8664 / 17544	8	non-labelled	III, IV, V
Dim32	1024	32	16	V
Sensor data - Smart Logistics			non-labelled	VI, VII
HAR data				VI, VII
Assets data	265	24	non-labelled	VIII

## 3.3 Research Methodology

Studies conducted in this thesis use research methodologies implementation, experimentation and case study [41]. Novel clustering approaches are proposed and evaluated using experimentation in each of the studies included in this thesis, except for paper IV. In paper IV case-study is used and the algorithm proposed in III is applied and evaluated in the smart building domain.

Various types of experiments are conducted to validate the algorithms using more data sets (more information about the data sets used is present in Section 3.2). In Paper I, the proposed algorithm, *Split-Merge Evolutionary Clustering* has been

compared with two other state-of-the-art algorithms, namely *PivotBiCluster* [26] and *Dynamic Split and Merge Clustering* algorithm [11] using the cover-type and wine quality data sets. Along with comparing to the state-of-the-art algorithms, experimentation was also done to evaluate if the size of newly arriving data has any impact on the algorithm’s performance. Yeast and anthropometric data sets are used for this purpose.

Papers II and III propose novel multi-view clustering approaches, entitled *MV Split-Merge Clustering* and *MV Multi-Instance Clustering*, respectively. Paper II is an initial study where the proposed algorithm is evaluated on anthropometric data. The proposed algorithm is also compared to its batch version. Paper III tries to improve the performance, understandability and interpretability of the results in comparison to the algorithm proposed in Paper II and has been evaluated on both anthropometric and real-world sensor data. The results obtained on the anthropometric data are compared to the results obtained in Paper II.

Paper IV is an applied work where the potential of *MV Multi-Instance Clustering* is investigated in the smart building domain. Real-world sensor data of the heating and tap-water systems from the smart building systems is used in the study. Two different sets of experiments are designed and performed to showcase the algorithm’s potential in context aware modeling of system behaviour and Integration analysis of system performance.

In Paper V, the proposed MST based multi-view algorithm (*MST-MVS*) has been evaluated on three data sets (Dim32, Cover-type, and Real-world sensor data) and has been compared to *MV Multi-Instance Clustering* (algorithm proposed in Paper III) on the real-world sensor data set. Two new approaches, BNodes and LEdges, are proposed to calculate the artificial nodes that is used to transfer the knowledge extracted from the global model to be used to cluster the newly arriving data chunks. Various experiments are conducted to evaluate and justify that different steps included in the algorithm improve the quality of the final clustering solution generated. In one experimental setup, the usefulness of the knowledge transfer is studied. Another set of experiments are done to understand how the quality of the local models affects the built global model.

Papers VI and VII propose unsupervised domain adaptation algorithms, *DIBCA* and *DIBCA++*. Both these papers use Sensor data from smart logistics domain from our industrial partner. Paper VI uses a publicly available HAR data set (HAR-1) to showcase the algorithm’s potential in the automatic data labeling task and Sensor data obtained from one device but at two different locations, is used to showcase the potential in the domain adaptation task. Paper VII also uses HAR data set (HAR-2) and Sensor data obtained from five different devices. Experiments are conducted to evaluate the performance with respect to the domain adaptation task and also understand how the similarity between data in different domains effects *DIBCA++*. *DIBCA++* is also experimentally compared to its predecessor *DIBCA*.

In paper V, a ...



hybrid clustering methodology, combining multi-layer data analysis with shared nearest neighbor similarity (SNNS) and hypergraph clustering, that can be used to interpret and analyze heterogeneous multi-source data with missing values

### 3.4 Evaluation Metrics

Different types of evaluation metrics have been used across the papers included in the thesis. Clustering models can be evaluated using either internal or external validation metrics [42]. External validation metrics use external information not used to build the clustering model such as the labels to validate the model. Whereas it is the opposite for internal validation metrics, they use the same information that is used to build the clustering solutions. Internal measures assess compactness, separation, connectedness, and stability aspects of clustering solutions. External validation metrics are further divided into unary and binary [43], and the third category, information theory [44] is also used by some authors. Table 3.2 gives an overview of the types of clustering metrics used in different papers included in the thesis.

Apart from evaluating the final clustering solutions, the Silhouette Index (SI) is also used in the clustering process. In Papers I and III, SI is used to find the optimal number of clusters when using  $k$ -means to do the initial clustering of the data points in a chunk. In Paper IV it is used to evaluate the quality of the local clustering solutions to decide which ones to be used for building the global model.

**Table 3.2:** Evaluation metrics used across papers

Evaluation metrics	Type of metric	Papers
Silhouette Index	Internal measure	I
F-measure	Unary	I, VI
Jaccard Index	Binary	I
Adjusted Rand Index	Binary	II, III, V, VII
Adjusted Mutual Information	Information theory	V, VII
Purity/Homogeneity	Information theory	II, III, V
Completeness	Information theory	V
Accuracy		VI

### 3.5 Validity Threats

This section describes the various validity threats that could have occurred in the process of conducting the thesis and the measures taken to overcome these. In general, all the papers have undergone a peer-review process and are published (except Paper VII, which is currently under review) in various conferences and Journals.

### 3.5.1 Internal Validity Threat

Internal validity addresses the concerns related to the effects of experimental setup on the results [45, 46]. To avoid selection bias while dividing the data set into different chunks, 10 different test sets of the data set are used for the conducted experiments in all experimental scenarios except when real-world sensor data is used, and the average value (Paper IV uses the minimum and maximum values as well) is considered as the result. In addition, devoted experiments are conducted to select the algorithm's parameters in most cases. E.g., In Paper IV, we studied different sizes of data chunks, different techniques for finding artificial nodes, etc.; the elbow method based on SI score is used to determine the  $k$  value when using the  $k$ -means algorithm.

### 3.5.2 External Validity Threat

External validity reflects the generalisability of the results obtained in a study [45, 46]. Experiments conducted across all the papers included are designed with caution to mitigate such threats. More than one data sets are used to evaluate the proposed algorithm in all papers except in Paper II, to avoid results specific to a particular scenario or use case. As Paper II is an initial study only a single data set is used. Testing the algorithm on more than one data set can generalize the performance of the algorithm when compared to using it just on one data set.

### 3.5.3 Construct Validity Threat

Construct validity addresses the concerns related to the results being deviated from the desired conceptual output [45, 46]. Such threats could become a reality if the implemented algorithm does not produce results similar to the algorithm designed during the design phase. To avoid such threats, the proposed algorithms, experimental scenarios, and setups are well discussed within the research group before the actual implementation phase has begun. During the implementation phase care is taken by conducting periodical tests to make sure that the functionality of the code is as expected. This is done to avoid logical and/or run time errors which are difficult to identify, compared to the compile time errors.

# 4 Results and Analysis

This chapter summarizes the overall results of the thesis. Finally, all the research questions formulated in this thesis are stated and answered.

In Paper I, a novel *Split-Merge Evolutionary clustering* algorithm based on bipartite correlation clustering is proposed. In the conducted experiments, the proposed algorithm has exhibited its ability to integrate newly arriving data continuously. When compared to two other state-of-the-art algorithms, *PivotBiCluster* [26], and *Dynamic split and merge clustering* [11], it is observed that different algorithms excelled for different evaluation metrics. This validates what Luxburg et al. [47] has stated that different cluster validation metrics can sometimes produce contradictory results, and it is difficult to validate how the clustering algorithm performs using these validation metrics. The author, therefore, highlights the importance of studying clustering algorithms considering their final usage and not as an application independent problem. It is interesting to discover that the number of clusters produced by the proposed *Split-Merge Evolutionary clustering* is closer to the benchmark solution when compared to the other two algorithms.

Multi-view stream clustering algorithms proposed in Papers II and III are capable of performing horizontal as well as vertical data integration. That is, the algorithms are capable of integrating the clustering solution of newly arriving data chunks with the existing clustering solution, and they can integrate the knowledge from different views into a global model. Note that the algorithms provide flexibility in choosing which views to be considered for building the global model. If three views are available, the algorithms can use all three views or just two of these views to build the global model. This type of functionality can be advantageous if information from any view is missing due to system or device failures or other similar reasons. The *MV Split-Merge clustering* algorithm proposed in Paper II when compared to its batch version has produced comparable results. The algorithm proposed in Paper II is outperformed by *MV Multi-Instance clustering* algorithm proposed in Paper III for the experiments conducted on the anthropometric data set. In addition, when evaluated on the real-world sensor data, the *MV Multi-Instance clustering* algorithm (Paper III) was able to perform continuous monitoring, analysis, and mining of streaming data. When new data chunks arrive, the algorithm is able to perform stable integration. It is observed that some concepts from the previous chunk are retained or expanded with new instances. The algorithm has also successfully detected devi-

ating behavior already known to have occurred in the system during the considered time.

*MST-MVS* algorithm proposed in Paper IV has performed well on the synthetic data (Dim32) compared to other data sets considered. This behavior can be backed logically as the synthetic data is free from factors such as noise, outliers, and cluster overlap that could exist in real-world data sets. Note that the Cover-type data set obtained from the UCI ML repository is also a real-world data set. The results indicate that the transfer of knowledge from the global model to build the local clustering solutions of next data chunk is beneficial in the considered experimental scenarios. That is, the quality of the local clustering solutions generated using the transferred knowledge is better than the one where this knowledge is not available. Apart from these, the study has also proposed a post-labeling technique used to label the data points of each view into different clusters obtained in the global model. This labeling technique has produced better results when compared to the Convex Non-negative Matrix Factorization (CNMF) based labeling technique.

## 4.1 Research Questions' Answers

**RQ1** *How can a clustering solution be updated to accommodate and catch evolving characteristics of continuously arriving data?*

Based on the research done and results obtained, it is observed that techniques like MST, bipartite graphs, and MI clustering can be used to integrate newly arriving data and update the existing clustering solution. According to Bahri et al. [48], the algorithms proposed in Papers I, II, and III, can be categorised as the sliding window modeling and the *MV-MST* (Paper IV) as a landmark window model. Note that the proposed algorithms based on sliding window modeling do not have a hard restriction on the chunk sizes and they may vary, but algorithm always uses two chunks in each iteration (chunks  $t$  and  $t - 1$ ).

In Papers, I, II, and III, the newly arriving data chunk is initially clustered. Then the existing clustering model is updated based on the newly arriving data by taking into account the correlations between the two clustering solutions. Therefore, the updated clustering solution is based on the data characteristics of both the existing clustering solution and newly arriving data. Unlike the other papers, the algorithm proposed in Paper V does not update the existing clustering solution when a new data chunk arrives. Instead, a new clustering solution is built for each data chunk based on the current data (landmark window model) and knowledge from the previous data chunk in the form of the artificial node used in the MST.

**RQ2** *How can clustering models generated in a multi-view context be integrated into a robust global model capturing knowledge from different views?*

In Papers II and III of this study, FCA integrates the clustering models from different views and builds a consensus clustering model (global model). The global

model consists of a formal context and a concept lattice. While building the formal context, the data points are considered as the objects, and the cluster each data point belongs to in each view is represented as the object's properties. A concept lattice is a hierarchical structure of concepts, which present the relationships between the objects and its properties. All the data points (objects) in a concept belong to that concept's clusters (properties) and display the relationship between the clusters of different views. The number of instances in a concept represents how strong the relationship is between different clusters in that concept. In the conducted experiments, it is noticed that the concept lattice generated by the approach proposed in Paper II is large and complex due to the vast amount of data. The results also demonstrate that using closed patterns in Paper III has decreased the complexity of the concept lattice generated, increasing the understandability of the results obtained.

A novel integration approach has been presented in Paper V; the global model is obtained by using MST clustering on an integrated matrix. The integrated matrix is built using the representatives of each cluster from the local clustering models and their attributes from other views. This approach was able to successfully integrate information from different views, and build a global model. The results demonstrate that evaluating the quality of the local clustering models before using them to build the global model produces better clustering solutions. Even though the results are not significantly different, the scenario in which the local clustering solutions are evaluated produced better results when compared to not evaluating the local models with respect to all the considered metrics except for *minimum value of completeness*.

**RQ3** *How can we design a resource-efficient domain adaptation algorithm for a robust clustering model alignment to new domains?*

**RQ4** *What clustering analyses can be performed on heterogeneous multi-source data with missing values to produce useful homogeneous clusters?*



## 5 Conclusions and Future Work

This thesis has proposed and evaluated different clustering algorithms suitable for analyzing and mining evolving and heterogeneous data. Through the various phases (different papers) of the thesis, we have worked on improving the algorithms' robustness by dealing with new challenges that were not addressed in the earlier versions. The initial study (Paper I) focuses on developing and evaluating a single source evolutionary clustering algorithm entitled *Split-Merge Evolutionary Clustering* based on bipartite correlation clustering.

This is followed by three studies focusing on evolutionary clustering algorithms for mining and analysis of multi-view streaming data. The *Split-Merge Evolutionary Clustering* proposed in Paper I has been compared with two other state-of-the-art clustering algorithms. The results show that the proposed algorithm is capable of integrating newly arriving data chunks into the existing clustering solution. It was not easy to point out the best-performing algorithm among the considered algorithms; different algorithms showed different performance on different cluster validation metrics. These results imply the need to study and evaluate clustering solutions based on the context where they are useful rather than using different evaluation metrics as stated by Luxburg et al. [47].

Paper II proposes a novel multi-view clustering approach, *MV Split-Merge Clustering*, an extended version of *Split-Merge Evolutionary Clustering* that can handle multi-view streaming data. The clustering model and relationships between different views generated by *MV Split-Merge Clustering* are comparable to those obtained in the batch data scenario for the considered experimental scenarios.

The *MV Multi-Instance Clustering* algorithm (Paper III) has been enhanced compared to the algorithm proposed in Paper II by using new data mining techniques to obtain better performance and interpretability of the results. Initially, MI clustering, which can handle the ambiguity of real-world scenarios, is used to obtain the local clustering models at each view. Then, closed patterns, which contain frequently occurring patterns, are used to build the global model. Extracting closed patterns reduces the complexity of the concept lattice and makes it easy to interpret and analyze the results. In the considered experimental scenarios, *MV Multi-Instance Clustering* was able to detect deviating behaviors successfully and has also shown better perfor-

mance than *MV Split-Merge Clustering*.

Finally, in Paper IV, we have proposed a graph-based clustering using MST for multi-view streaming data, entitled *MST-MVS*. Unlike other algorithms proposed in this thesis, *MST-MVS* uses the knowledge transferred from the global model of previous chunk to build the local clustering solutions of the current chunk. This knowledge transfer has positively impacted the generation of local clustering solutions in the conducted experiments. The study also proposes a post-labeling technique to label the chunk's data points based on the obtained global model. The proposed labeling technique produced better results when compared to the CNMF based labeling technique. The algorithm is evaluated on both the real-world and synthetic data sets. The algorithm has demonstrated better performance on synthetic data compared to real-world data. We believe it is due to different types of challenges such as noise, complex cluster structure, etc., which come into play when dealing with real-world data.

Future work ...



## 6 Experiences and Learning Outcomes

The Ph.D. journey has been a great learning activity and has provided a lot of new experiences. It has helped me to grow as a researcher over time. During this time I got an opportunity to work on different interesting research problems, which are also very relevant for different application scenarios in the industry.

I acted as a co-supervisor for a master's thesis student, which has led to a successful publication (Paper V). It was a different experience being on the other side and added a new responsibility in guiding the student to conduct the research.

One of the major fears that I partially overcame in this journey is public speaking. Over these last few years, there have been numerous occasions like internal monthly seminars within our research group, workshops, and conferences where I had to present my research work. These events have helped me to a certain extent to overcome my fears. Writing was also not one of my strengths, but I have come a long way in this as well.



# Bibliography

- [1] A. Bifet, B. Hammer, and F. Schleif. “Recent trends in streaming data analysis, concept drift and analysis of dynamic data sets”. In: *27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, April 24-26, 2019*. 2019.
- [2] D. Xu and Y. Tian. “A Comprehensive Survey of Clustering Algorithms”. In: *Annals of Data Science* 2 (Aug. 2015), pp. 165–193. DOI: 10.1007/s40745-015-0040-1.
- [3] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. English. Englewood Cliffs, NJ: Prentice Hall, 1988, pp. xiv + 320. ISBN: 0-13-022278-X.
- [4] B. Jiang and et al. “Evolutionary multi-objective optimization for multi-view clustering”. In: *2016 IEEE CEC 2016*. 2016, pp. 3308–3315.
- [5] Y. Yang and H. Wang. “Multi-view clustering: A survey”. In: *Big Data Mining and Analytics* 1.2 (June 2018), pp. 83–107. ISSN: 2096-0654.
- [6] L. Fu, P. Lin, A. V. Vasilakos, and S. Wang. “An overview of recent multi-view clustering”. In: *Neurocomputing* 402 (2020), pp. 148–161. ISSN: 0925-2312.
- [7] A. Zubaroğlu and V. Atalay. “Data stream clustering: a review”. In: *Artificial Intelligence Review* 54.2 (2021), pp. 1201–1236. DOI: 10.1007/s10462-020-09874-x.
- [8] L. Huang and et al. “MVStream: Multiview Data Stream Clustering”. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.9 (2020), pp. 3482–3496.
- [9] W. Shao and et al. “Online multi-view clustering with incomplete views”. In: *2016 IEEE Int. Conf. on Big Data (Big Data)*. 2016, pp. 1012–1017.
- [10] A. Bouchachia. “Evolving clustering: an asset for evolving systems”. In: *IEEE SMC Newsletters* 36 (2011).
- [11] E. Lughofer. “A dynamic split-and-merge approach for evolving cluster models”. In: *Evolving Systems* 3.3 (Sept. 2012), pp. 135–151.
- [12] R. Fa and A. K. Nandi. “Smart: Novel self splitting-merging clustering algorithm”. In: *European Signal Processing Conference, Bucharest, Romania, August, 27-32*. IEEE, 2012.

- [13] M. Wang, V. Huang, and A.-M. C. Bosneag. “A Novel Split-Merge-Evolve k Clustering Algorithm”. In: *IEEE 4th International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, March 26-29. 2018.*
- [14] A. S. Iwashita and J. P. Papa. “An Overview on Concept Drift Learning”. In: *IEEE Access* 7 (2019), pp. 1532–1547. DOI: 10 . 1109 / ACCESS . 2018 . 2886026.
- [15] K. Wadewale and S. Desai. “Survey on Method of Drift Detection and Classification for time varying data set”. In: *Int. Res. J. Eng. Technol.* Vol. 2. 2015, pp. 709–713.
- [16] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. d. Carvalho, and J. Gama. “Data Stream Clustering: A Survey”. In: *ACM Comput. Surv.* 46.1 (July 2013). ISSN: 0360-0300. DOI: 10 . 1145 / 2522968 . 2522981. URL: <https://doi-org.miman.bib.bth.se/10.1145/2522968.2522981>.
- [17] M. Ghesmoune, M. Lebbah, and H. Azzag. “State-of-the-art on clustering data streams”. In: *Big Data Analytics* 1.1 (2016), pp. 1–27.
- [18] B. Cao, D. Shen, J.-T. Sun, X. Wang, Q. Yang, and Z. Chen. “Detect and Track Latent Factors with Online Nonnegative Matrix Factorization.” In: *IJCAI*. Vol. 7. 2007, pp. 2689–2694.
- [19] C.-D. Wang, J.-H. Lai, D. Huang, and W.-S. Zheng. “SVStream: A support vector-based algorithm for clustering data streams”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.6 (2011), pp. 1410–1424.
- [20] S. Huang, Z. Xu, I. W. Tsang, and Z. Kang. “Auto-weighted multi-view co-clustering with bipartite graphs”. In: *Information Sciences* 512 (2020), pp. 18–30. ISSN: 0020-0255. DOI: <https://doi.org/10.1016/j.ins.2019.09.079>.
- [21] M. Bendecheache and M.-T. Kechadi. “Distributed clustering algorithm for spatial data mining”. In: *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*. IEEE. 2015, pp. 60–65.
- [22] X. Liu and et al. “Late Fusion Incomplete Multi-View Clustering”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 41.10 (2019), pp. 2410–2423.
- [23] Y. Ye and et al. “Incomplete Multiview Clustering via Late Fusion”. In: *Computational Intelligence and Neuroscience* 2018 (Oct. 2018), pp. 1–11.
- [24] S. Wang and et al. “Multi-view Clustering via Late Fusion Alignment Maximization”. In: *Proceedings of IJCAI-19*. July 2019, pp. 3778–3784.

- [25] C. Zhu. “Kappa Based Weighted Multi-View Clustering with Feature Selection”. In: *Proceedings of ICCPR 2018*. ICCPR '18. Shenzhen, China, 2018, pp. 50–54. ISBN: 978-1-4503-6471-3.
- [26] N. Ailon, N. Avigdor-Elgrabli, E. Liberty, and A. van Zuylen. “Improved Approximation Algorithms for Bipartite Correlation Clustering”. In: *Algorithms - ESA 2011 - 19th Annual European Symposium, Saarbrücken, Germany, September 5-9, 2011. Proceedings*. 2011, pp. 25–36.
- [27] G. W. Flake, R. E. Tarjan, and K. Tsioutsouloukakis. “Graph clustering and minimum cut trees”. In: *Internet Mathematics* 1.4 (2004), pp. 385–408.
- [28] R. Görke, T. Hartmann, and D. Wagner. “Dynamic Graph Clustering Using Minimum-Cut Trees”. In: *Algorithms and Data Structures*. Ed. by F. Dehne, M. Gavrilova, J.-R. Sack, and C. D. Tóth. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 339–350.
- [29] B. Saha and P. Mitra. “Dynamic Algorithm for Graph Clustering Using Minimum Cut Tree”. In: *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*. 2006, pp. 667–671.
- [30] J. Foulds and E. Frank. “A review of multi-instance learning assumptions”. In: *Knowledge Engineering Review* 25.1 (2010), pp. 1–25. DOI: 10 . 1017 / S026988890999035X.
- [31] M. Zhang and Z. Zhou. “Multi-instance clustering with applications to multi-instance prediction”. In: *Applied Intelligence* 31 (2009), pp. 47–68.
- [32] B. Ganter, G. Stumme, and R. Wille. “Formal Concept Analysis: Foundations and Applications”. In: LNAI, no. 3626, Springer-Verlag, 2005.
- [33] V. Boeva and et al. “Analysis of Multiple DNA Microarray Datasets”. In: *Springer Handbook of Bio-/Neuroinformatics*. Ed. by N. Kasabov. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 223–234. ISBN: 978-3-642-30574-0.
- [34] A. Hristoskova, V. Boeva, and E. Tsiporkova. “A Formal Concept Analysis Approach to Consensus Clustering of Multi-Experiment Expression Data”. In: *BMC Bioinformatics* 15 (May 2014), p. 151.
- [35] S. K. and A. K. Ch. “Concept Lattice Simplification in Formal Concept Analysis Using Attribute Clustering”. In: *Journal of Ambient Intelligence and Humanized Computing* 10 (2018), pp. 2327–2343. ISSN: 1868-5145.
- [36] J. A. Blackard, D. J. Dean, and C. W. Anderson. *UCI Machine Learning Repository*. 1998. URL: <http://archive.ics.uci.edu/ml>.
- [37] K. Nakai and M. Kanehisa. “Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria”. In: *PROTEINS: Structure, Function, and Genetics* 11 (1991), pp. 95–110.

- [38] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reisa. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009), pp. 547–553.
- [39] H. F. Golino, L. S. de Brito Amaral, S. F. P. Duarte, and et al. “Predicting Increased Blood Pressure Using Machine Learning”. In: *Journal of Obesity* 2014 (2014).
- [40] P. Fränti and S. Sieranoja. *K-means properties on six clustering benchmark datasets*. 2018. URL: <http://cs.uef.fi/sipu/datasets/>.
- [41] “Developing your Objectives and Choosing Methods”. In: *Thesis Projects: A Guide for Students in Computer Science and Information Systems*. London: Springer London, 2008, pp. 54–70. ISBN: 978-1-84800-009-4. DOI: 10.1007/978-1-84800-009-4\_8. URL: [https://doi.org/10.1007/978-1-84800-009-4\\_8](https://doi.org/10.1007/978-1-84800-009-4_8).
- [42] K. A. Jain and C. R. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.
- [43] J. Handl, J. Knowles, and D. Kell. “Computational cluster validation in post-genomic data analysis”. In: *Bioinformatics* 21.15 (2005), pp. 3201–3212.
- [44] H. Van der Hoef and M. J. Warrens. “Understanding information theoretic measures for comparing clusterings”. In: *Behaviormetrika* 46 (2019), pp. 353–370.
- [45] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, 2002.
- [46] R. Feldt and A. Magazinius. “Validity Threats in Empirical Software Engineering Research - An Initial Survey”. In: *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE'2010), Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010*. Knowledge Systems Institute Graduate School, 2010, pp. 374–379.
- [47] U. von Luxburg, R. C. Williamson, and I. Guyon. “Clustering: Science or Art?” In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. Vol. 27. Proceedings of Machine Learning Research. 2012, pp. 65–79.
- [48] M. Bahri, A. Bifet, J. Gama, H. Gomes, and S. Maniu. “Data stream analysis: Foundations, major tasks and tools”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.3 (2021). DOI: 10.1002/widm.1405.