

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 May 22nd, 2024, at 09:00 in room J1630, Campus Gräsvik

Supervisors

Prof. Veselka Boeva, Blekinge Institute of Technology, Sweden
Prof. Niklas Lavesson, Blekinge Institute of Technology, Sweden

Faculty Opponent

Asst. Prof. Shehroz Khan, KIET, Toronto Rehabilitation Institute, Canada

Grading Committee

Prof. Ivan Koychev, Sofia University, Bulgaria
Assoc. Prof. Sindri Magnusson, Stockholm University, Sweden
Assoc. Prof. Farhana Zulkernine, Queen's University, Canada

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 evolving 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, Evolving clustering, Domain adaptation, 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

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 research publications related to but not included in the thesis are:

Paper IX

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 X

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 XI

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 the master thesis project at the foundation of this study.

Abbreviations

FCA	Formal Concept Analysis.
HAR	human activity recognition.
IoT	Internet of Things.
MI	multi instance.
ML	Machine Learning.
MST	minimum spanning tree.
SI	silhouette index.
SNNS	shared nearest neighbor similarity.

Table of Contents

Acknowledgements	i
List of Papers	iii
Abbreviations	vii
Chapter 1 Introduction	1
1.1 Research Problem	2
1.2 Contributions and Papers Included	5
1.3 Thesis Structure	7
Chapter 2 Background and Related Work	9
2.1 Evolving Clustering	9
2.2 Stream Clustering	10
2.3 Multi-View (Stream) Clustering	10
2.4 Domain Adaptation	11
2.5 Graph-based Clustering	11
2.6 Multi-Instance Clustering	12
2.7 Formal Concept Analysis	12
Chapter 3 Methodology	13
3.1 Data sets	13
3.2 Evaluation measures	13
3.3 Research Methodology	15
3.4 Validity Threats	16
3.4.1 Internal Validity Threat	16
3.4.2 External Validity Threat	16
3.4.3 Construct Validity Threat	17
3.4.4 Conclusion Validity Threat	17
Chapter 4 Results and Analysis	19
4.1 Evolving Clustering	19
4.2 Multi-Source Data Analysis	21
4.3 Domain Adaptation	25
4.4 Summary	27
Chapter 5 Conclusions and Future Work	29
Chapter 6 Experiences and Learning Outcomes	31

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 categories. 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. In addition to this, they are capable of mining hidden information from 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 focuses 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 causing the occurrence of phenomena like concept drift [4] or domain shift. It is an important aspect to be considered while developing algorithms suitable for evolving data. When the data characteristics of newly arriving data are no longer the same, it becomes difficult to accommodate them into the existing model as it becomes outdated [5]. This provides the need for evolving algorithms, which can be updated regularly to be suitable for new data. Many computational resources are required to rebuild the model when new information is generated. Hence, evolving 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 vari-

ous 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 [6]. 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 [7].

Traditional clustering algorithms are unsuitable for addressing the above-stated challenges of multi-source [8] and streaming [4] data. In this thesis, novel clustering approaches are developed 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. In this process, three evolving clustering algorithms, four multi-view clustering analysis algorithms/approaches, and two domain adaptation algorithms are proposed. Eight different studies have been conducted as a part of this PhD thesis. This thesis aims to develop clustering techniques to mine and analyze streaming data by handling its evolving, heterogeneous, and multi-source nature.

- Obj 1. To propose data analysis 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 propose clustering-based analysis methods that can handle data with missing values generated from multiple heterogeneous sources.

Based on the aims and objectives set to be accomplished, the following research questions are framed and addressed 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: Many current-day applications continuously generate a lot of data whose characteristics tend to change over time. It becomes difficult to accommodate and fit new data into the current model in such contexts. In such situations, one alternative is to rebuild the clustering model, which is not optimal as (i) it would consume a lot

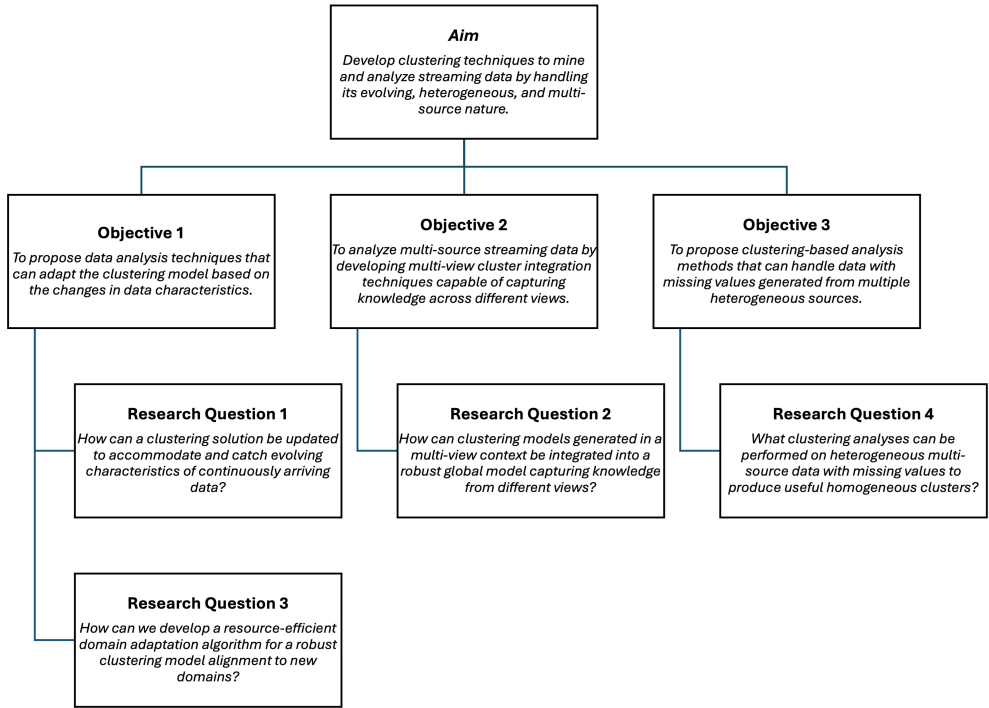


Figure 1.1: Figure visualizing the connections between the thesis's aim, objectives, and research questions.

of computational resources, and (ii) in such situations, importance should be given to the latest trends while retaining the existing knowledge which would not be completely possible by rebuilding the model from scratch. Hence, there is a need for a clustering approach that can integrate new data into the existing model based on its characteristics.

Papers: Paper I proposes a novel *Split-Merge Evolutionary clustering* technique, also used in Paper II. The algorithm can split or merge existing clusters based on the clustering model of newly arriving data, thus obtaining one final updated clustering model. In Paper III, *Bi-Correlation MI-clustering* is proposed based on multi instance (MI) learning and bipartite correlation clustering. This algorithm is evaluated in a real-world use case of smart building systems in Paper IV. Another algorithm *MST-MVS* proposed in Paper V.

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

Motivation: Multi-source data is common in many application scenarios, and information from these sources could complement the other. When viewed together, this data can give or produce information that cannot be obtained when only each of these views is considered [6]. 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 [9, 10].

Papers: This research question is addressed in papers II, III, IV, and V, where the first three papers use Formal Concept Analysis (FCA) to integrate knowledge from different views. In addition to FCA, closed patterns were also used in papers III and IV to extract frequent patterns and reduce the complexity of the global model. In Paper V, the local clustering models of different views are initially evaluated to avoid negative impact and those of the desired quality (silhouette index (SI) greater than the set threshold) are used to build the global model. The selected representatives of the cluster and their attributes from other views are extracted to build an integrated matrix, which is clustered using a minimum spanning tree (MST) based clustering algorithm.

RQ3 *How can we develop 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 use 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 edge devices with strict resource constraints. Novel domain adaptation techniques that are resource-efficient and can support robust model adaptation to new contexts are needed.

Papers: Papers VI and VII address this research question by proposing novel unsupervised domain adaptation algorithms, namely *DIBCA* and *DIBCA++*, which are based on cluster integration. 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 developed to be resource efficient as all the operations are performed using only the representatives.

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

Motivation: In many application scenarios related to monitoring the performance of different types of assets, data is usually generated from multiple sources and is likely to be heterogeneous. Analyzing and deriving meaningful full insights from such data is challenging [11]. As stated before, certain interesting aspects might be associated with a subset of features. In such cases, using multi-view or multi-layered techniques to analyze a few features at a time might be helpful [12–14]. Another common concern with data from multiple sources is having many missing values due to lack of standardization, equipment malfunctioning, registration errors, communication issues, etc. Common practices to deal with missing values include imputation techniques or removing the features with a high degree of missing values, but such practices negatively affect the data quality [15]. This raises the need for studying

and proposing clustering techniques suitable to analyze multi-source heterogeneous data with missing values. The obtained homogeneous groups are expected to have similar behavior and are useful for tasks like performance monitoring.

Papers: Paper VIII mainly focuses on this research question. A multi-layered clustering approach followed by hypergraph clustering based on k-medoids and nearest-neighbor similarity is proposed to achieve this. The other multi-view clustering algorithms proposed in this thesis are also capable of handling heterogeneous data. However, *MST-MVS* is not capable of handling missing values and it can be noted that even if the others are capable of working in scenarios with missing values, they are not evaluated concerning this aspect in those studies.

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 multi-source challenges, and all the papers except Paper VIII propose or use algorithms to address challenges related to the evolving nature and domain shift (when multiple devices or locations are considered) of the streaming data. 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 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 developed to analyze multi-view streaming data and build a consensus clustering solution (global model) based on the information obtained from different views where FCA is used for the integration. An initial evaluation is done, and the algorithm is compared to its batch version, where 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 developed 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. Closed patterns are also used to mine frequent concepts while building the global model, reducing 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 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 build the clustering solution in the next chunk by considering how the previous data has been grouped. A post-labeling technique is also proposed to label the chunk’s data points based on the built global model. *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 developed 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 algorithm’s explainability aspects and applica-

bility potential 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 developed 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 2 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 3, which offers an overview of the scientific methodology used in this thesis. It covers data sets used, evaluation measures, research methodology, and validity threats. The results of the thesis and their analysis are presented in Chapter 4, and the conclusions and future work are presented in Chapter 5. In Chapter 6, experiences and the learning outcomes identified during the PhD journey are presented.

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 Evolving Clustering

Evolving clustering algorithms are designed to continuously accommodate data with evolving characteristics, i.e., the data characteristics might change over time. 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 [16], different phases of an evolving clustering algorithm can be categorised into matching, accommodating new data, and model refinement.

In papers [17–19] 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 [17]. 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.

Concept drift is a phenomenon where the data characteristics tend to change over time [4]. Thereby making the ML model outdated and degrading its performance if concept drift is not handled [5]. 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 [20].

2.2 Stream Clustering

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 [4]. While designing stream clustering algorithms, it is essential that the time and memory constraints are considered for handling large amounts of data. Concept drift is a common problem that needs to be addressed by a clustering algorithm designed for streaming data [4].

In [4, 21, 22], the authors review the existing state-of-the-art stream clustering algorithms. An iterative Non-negative matrix factorization based algorithm is proposed in [23], entitled as ONMF. Another stream clustering algorithm, SVStream is proposed in [24], which is a predecessor of MVStream [9], a multi-view stream clustering algorithm.

2.3 Multi-View (Stream) 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 [6]. 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 [25].

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 [26]. Incomplete views, that is, the possibility of missing information from different views is another common challenge. Authors of [10, 27, 28] address the later challenge in their works.

Studies [7, 8] 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. [29] 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 [10]. Zhu et al. [30] have proposed another algorithm in the field that is based on feature selection.

Multi-view stream clustering algorithms are expected to address the challenges of both multi-view and stream clustering algorithms as the name suggests. 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 [9, 10]. Authors of [9, 10] have proposed novel

algorithms in the field. MVStream proposed in [9] is based on support vectors, and is robust to concept drift. In [10], online multi-view clustering based on non-negative matrix factorization is proposed.

2.4 Domain Adaptation

Machine learning models trained in one domain (location, device, etc.) do not always perform well when used in another domain due to changes in data characteristics. In such situations, the existing model needs to be adapted to the newer data characteristics. The domain where the model is initially trained is referred to as the source, and the one to which this model is adapted to is referred to as the target [31]. Domain adaptation is a sub-branch of transfer learning, where the source and target address similar problems but have different data characteristics [32]. The quintessential ML rule that the train and test data have the same data characteristics does not hold in this case [33]. Source data privacy is a concern with many algorithms in the field, authors of [34, 35] propose source-free domain adaptation algorithms and state the importance of these.

Madadi et.al [36] presents a survey studying the unsupervised domain adaptation techniques that could be used for classification tasks. Various works have been done in the field of domain adaptation, but the majority of these works are in the fields of computer vision and deep learning. There haven't been many works addressing domain adaptation for the time-series data sets. In their work, the authors of [37] tested the performance of two of the existing domain adaptation algorithms on four different time series data sets. They have identified that the similarity of the source and target data sets impacts the performance. Li et al. [38], in their work, have proposed a novel domain consensus clustering algorithm to be able to handle source and target have different labels. The algorithm deals with this by separating the common and private clusters of the domains.

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 [39], 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. [39] 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 [40], 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 α . 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 [41, 42] 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 [43]. 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 [44].

2.7 Formal Concept Analysis

Formal Concept Analysis (FCA) [45] 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 [46–48] 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 adaptive and evolving clustering, multi-view clustering, and domain adaptation. This thesis aims to develop clustering techniques to mine and analyze streaming data by handling its evolving, heterogeneous, and multi-source nature. This chapter presents the data sets followed by the evaluation measures used in this thesis. Then, the research methodology is presented. Finally, the validity threats of the studies conducted are presented.

3.1 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. Ten different types of data sets are used across papers of this thesis namely, cover type [49], yeast [50], wine quality [51], anthropometric data [52], Dim32 [53], PAMAP2 [54], DaLiAc [55], assets data from our industrial partner, real-time sensor data from smart building and smart logistics domains. Three of these 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. Dim32 is a publicly available synthetic data set. PAMAP2 and DaLiAc are HAR data sets that are used to evaluate the domain adaptation algorithms. In addition to the publicly available data set, three data sets, assets data, real-time sensor data from smart building, and smart logistics domains, are obtained from our industrial partners. Table 3.1 presents the details about the data sets used in different papers, like number of attributes and classes available.

3.2 Evaluation measures

Different evaluation measures have been used across the papers in the thesis. Clustering models can be evaluated using either internal or external validation measures

Table 3.1: Data sets used in the thesis.

Data set	Attributes	Classes	Papers
Cover-type	14	7	I, V
Yeast	8	10	I
Wine quality	12	7	I
anthropometric data	9	6	I, II, III
Sensor data - Smart building Domain	8	not labelled	III, IV, V
Dim32	32	16	V
Sensor data - Smart Logistics	8	not labelled	VI, VII
PAMAP2 (HAR)	31	17	VI
DaLiAc (HAR)	24	13	VII
Assets data	24	not labelled	VIII

[56]. External validation measures 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 measures, 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 measures are further divided into unary and binary [57], and the third category, information theory [58] is also used by some authors. Table 3.2 gives an overview of the types of clustering measures 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, III, and IV, 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 VIII, SI along with Calinski Harabasz, Davies Bouldin, and Connectivity are used to determine the optimal number of clusters while using k -means to cluster the data in different layers (multi-layered clustering approach is used in this paper). In Paper V SI is used to evaluate the quality of the local clustering solutions to decide which ones are to be used for building the global model. It can be noted that in papers IV and VIII, the final results are evaluated using prior domain knowledge available with the help of experts and statistics.

Table 3.2: Evaluation measures used across papers

Evaluation measures	Papers
Internal	
Silhouette Index	I, VIII
Davies Bouldin	VIII
Calinski Harabasz	VIII
Connectivity	VIII
External	
Adjusted Rand Index	II, III, V, VII
F-measure	I, VI
Jaccard Index	I
Accuracy	VI
Information theory	
Adjusted Mutual Information	V, VII
Purity/Homogeneity	II, III, V
Completeness	V

3.3 Research Methodology

Studies conducted in this thesis use research methodologies implementation, experimentation, and case study [59]. 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, a 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 different types of data sets (more information about the data sets used is present in Section 3.1).

In Paper I, the proposed algorithm, *Split-Merge Evolutionary Clustering* has been compared with two other state-of-the-art algorithms, namely *PivotBiCluster* [39] and *Dynamic Split and Merge Clustering* algorithm [17] using the cover-type and wine quality data sets. Along with comparisons to the state-of-the-art algorithms, experimentation was also done to evaluate if the size of newly arriving data impacts the algorithm's performance, for which yeast and anthropometric data sets are used.

Papers II and III propose novel multi-view clustering approaches, entitled *MV Split-Merge Clustering* (based on *Split-Merge Evolutionary 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 compared to the algorithm proposed in Paper II and has been evaluated on both anthropometric and real-world sensor data of the heating system from the smart building domain. 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 are 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 behavior and integration analysis of system performance. The study also presents different visualization techniques to aid domain experts in understanding the results and using them to analyze and monitor 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 of the heating system from the smart building domain), it 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 are 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 is 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++*, respectively. Both these papers use sensor data from the 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 is used to showcase its potential in the domain adaptation task. The sensor data used in this paper is obtained from a single device operating at two different locations. Paper VII uses another HAR data set (HAR-2) and sensor data obtained from five different devices. The sensor data is used to understand how the similarity between data from different domains (devices) affects *DIBCA++*. Using both data sets the ability of *DIBCA++* to transfer new knowledge between the domains is showcased. *DIBCA++* is also experimentally compared to its predecessor *DIBCA* in Paper VII.

Paper VIII proposes a hybrid clustering methodology, combining multi-layer data analysis with shared nearest neighbor similarity (SNNS) and hypergraph clustering. The algorithm is developed such that it can be used to interpret and analyze heterogeneous multi-source data with missing values. A real-world data set related to condition monitoring of assets is used to validate the approach. The obtained homogeneous clusters are used to derive KPIs and are evaluated on the sensor data related to these compressors.

3.4 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.

3.4.1 Internal Validity Threat

Internal validity addresses the concerns related to the effects of experimental setup on the results [60–62]. To avoid selection bias while dividing the data set into different chunks, 10 different test sets of the data set are used for papers I, II, III, and V, the conducted experiments in all experimental scenarios except when real-world sensor data is used, and the average value (Paper V uses the minimum and maximum values as well) is considered as the result.

3.4.2 External Validity Threat

External validity reflects the generalisability of the results obtained in a study [60–62]. Experiments conducted across all the papers included are designed with caution to mitigate such threats. In most papers, more than one data set is used to evaluate the proposed algorithm to avoid results specific to a particular scenario or use case. Only papers II, IV, and VIII use one type of data set for evaluation. Paper II is an

initial study and hence only a single data set is used. Paper IV uses the algorithm proposed in Paper III (where initial evaluation has been done on other data sets) in an applied scenario of smart building systems. Paper VIII is also an applied study where the proposed approach is tested on the use case related to performance monitoring of industrial assets. 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. However, there might still be some use cases or scenarios where the proposed algorithms are not appropriate. Luxburg et al. [63] in their study highlights the need to study clustering algorithms as an application-dependent problem.

3.4.3 Construct Validity Threat

Construct validity addresses the concerns related to the results being deviated from the desired conceptual output [60–62]. Such threats could become a reality if the implemented algorithm does not produce results similar to the algorithm developed during the development phase. To avoid such threats, the proposed algorithms, experimental scenarios, and setups are well discussed within the research group before the implementation phase begins. During the implementation phase, care is taken by conducting periodical tests to ensure that the code’s functionality is as expected. This is done to avoid logical and/or run time errors, which are difficult to identify compared to compile time errors.

Devoted experiments are conducted to select the algorithm’s parameters in most cases. For Paper IV, we studied different sizes of data chunks, different techniques for finding artificial nodes, etc. Empirical methods have been used to determine the number of clusters when this information is unavailable. Papers I, III, and IV use the elbow method based on the SI score to determine the k value when using the k -means algorithm; Paper VIII uses other cluster evaluation measures along with SI as stated in Section 3.2.

3.4.4 Conclusion Validity Threat

Conclusion validity deals with the effectiveness of the research in terms of treatment of data, how the experiments are conducted, and the obtained outcomes [62]. 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 which validates the experimental treatment used across the papers.

4 Results and Analysis

This chapter summarizes the results of the thesis. We have identified three main research domains to which the thesis has contributed: evolve clustering, multi-source data analysis, and domain adaptation. The thesis achievements are presented and discussed with respect to these domains. Furthermore, potential application scenarios are outlined for the algorithms proposed in each domain. Finally, all the research questions formulated in this thesis are stated and answered.

4.1 Evolving Clustering

The results included in this thesis and contributing to the evolving clustering domain are presented in papers I, II, III, IV, and V. Papers I, III, and V propose novel evolving clustering algorithms, namely *Split-Merge Evolutionary clustering*, *Bi-correlation MI-clustering*, and *MST-MVS*. *Split-Merge Evolutionary clustering* is also used in paper II in a multi-view context. It can be noted that *Bi-correlation MI-clustering* is a part of *MV Multi-Instance Clustering* algorithm and *MST-MVS* are developed to handle multi-view data in a streaming fashion. Paper IV is an applied paper using the algorithm proposed in paper III in the smart building domain. Algorithms proposed/used in papers I, II, III, and IV use bipartite correlation clustering based on different approaches. 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 models. Based on the correlations, the existing clusters are either merged, split, or retained as they were, and a new, updated clustering model is obtained. Figure 4.1 (taken from Paper I) visualizes the bipartite correlation clustering or the split-merge framework of the *Split-Merge Evolutionary clustering* where C and C' represent the existing and new clustering models, respectively.

In the conducted experiments, *Split-Merge Evolutionary clustering* has exhibited its ability to integrate newly arriving data continuously. When compared to two other state-of-the-art algorithms, *PivotBiCluster* [39], and *Dynamic split and merge clustering* [17], it is observed that different algorithms excelled for different evaluation measures. This validates what Luxburg et al. [63] have stated, that different cluster validation measures can sometimes produce contradictory results, and it is

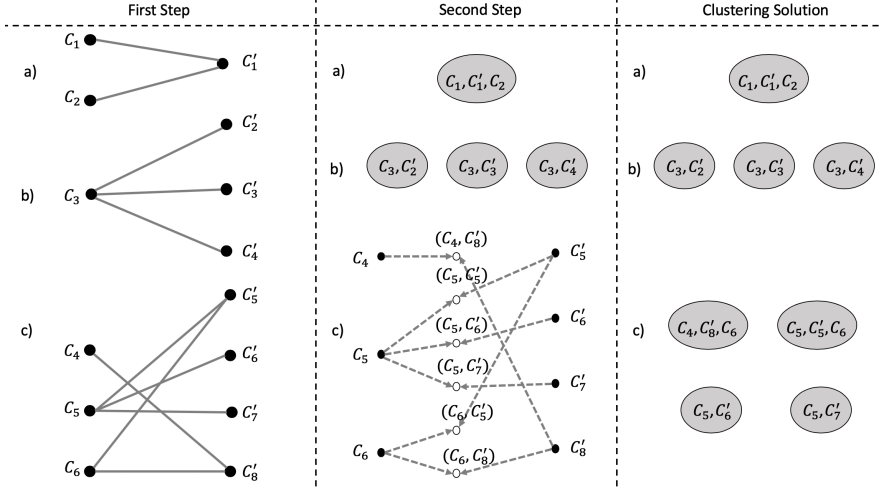


Figure 4.1: Visualization of Split-Merge framework in three different scenarios. a) bi-clique with under-clustered nodes (C_1 and C_2 correlate with C'_1); b) bi-clique with over-clustered nodes (C_3 correlates with C'_2 , C'_3 and C'_4); c) a bi-clique where there is a need to be decomposed into subcomponents, the second step presents the transformation into tripartite graph with split (left) and merge (right) subcomponents. The figure is taken from Paper I.

difficult to validate how the clustering algorithm performs using these validation measures. Luxburg et al., therefore, highlight 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.

Bi-correlation MI-clustering developed in paper III is based on MI clustering and bipartite correlation clustering. In *Split-Merge Evolutionary clustering*, only representatives from the clusters are 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). The experimental results also show improved performance when MI-Clustering is used (see Section 4.2). 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.

The algorithms proposed/used in Papers I, II, III, and IV can be categorized as the sliding window processing models and the *MV-MST* (Paper V) as a landmark window model based on the classification of processing methods of streaming data from the literature [64]. Note that, unlike regular sliding window models, where one or more elements are processed in each fixed-element counting window, the proposed algorithms categorized into this do not have a hard restriction on the chunk sizes, and

they may vary, but the algorithm always uses two chunks in each iteration (chunks t and $t - 1$).

Notice that the above-discussed algorithms automatically adapt the clustering solution to handle the changes in data characteristics due to phenomena like concept drift. We want to study this aspect in future works further to be able to analyze and understand how the proposed algorithms fair in different types of concept drift scenarios, classified as sudden, incremental, gradual, recurring, blip, or noise [20].

Application Scenarios: These proposed algorithms could be used in a wide range of real-world application scenarios where the data is generated in a continuous fashion and characterized by an evolving nature. Some examples include: (i) analysis and performance monitoring of heating sub-system in the smart building domain; due to changes in behavioral patterns of humans and climate changes, the normal system behavior is not constant and is evolving. (ii) patient profiling and monitoring; these models could be used for precision medicine, where patients are divided into groups to provide personalized treatment. Patients in a group are expected to have similar characteristics, thus requiring similar treatment. As more information from different patients is obtained, the existing clustering model could become outdated due to factors like advancements in the medical fields, the arrival of patients with new disease characteristics, etc., and the need arises to recluster the existing clustering model to be able to adapt to the new trends, where the proposed evolving clustering algorithms could be used. Papers III, IV, and V have evaluated the functionality of the proposed algorithms in the smart building domains.

4.2 Multi-Source Data Analysis

Papers II, III, IV, V, and VIII of this thesis contribute to the area of multi-source data analysis. Novel multi-view clustering algorithms *MV Split-Merge clustering*, *MV Multi-Instance clustering*, and *MST-MVS* are proposed in papers II, III, and V, respectively. As stated before, paper IV is an applied work where *MV Multi-Instance clustering* is evaluated in the smart building domain. Its potential in analyzing and monitoring the sub-systems (heating and tap-water systems are considered in the study) of the smart building domain is demonstrated. Paper VIII presents a hybrid clustering workflow for analyzing incomplete heterogeneous data.

FCA is used in papers II, III, and IV to integrate 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. Multi-view stream clustering algorithms proposed in Papers II and III can perform both horizontal and vertical data integration. That is, the algorithms can integrate 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. These algorithms also provide flexibility in choosing which views to consider 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. Figure 4.2 visualizes a high-level overview of the functionality of the *MV Split-Merge clustering* and *MV Multi-Instance clustering* algorithms with three different views.

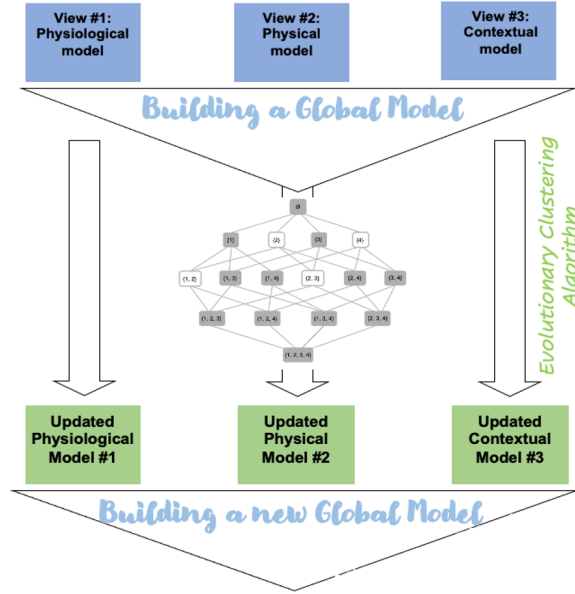


Figure 4.2: High-level overview of the functionality of *MV Split-Merge clustering* and *MV Multi-Instance clustering* algorithms where three different views are identified.

When compared to its batch version, the *MV Split-Merge clustering* algorithm proposed in Paper II 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 deviating behavior that is already known to have occurred in the system during the considered time.

Paper IV is an applied paper where *MV Multi-Instance clustering* is evaluated in the smart building domain. The heating and tap-water sub-systems, which depend on

each other, are studied in different scenarios. The study demonstrates the algorithm’s potential in performing context-aware modeling and integrated system analysis for modeling and analyzing system behaviors. Various visualization and data mining techniques are proposed for each step of the algorithm, which can aid experts in step-by-step analysis and easy understandability of the results. The algorithm successfully detected some known and unknown deviating behaviors that have occurred in the system during the time considered for experimentation. The experimental results show the ability of *MV Multi-Instance clustering* to monitor, analyze, and detect deviating behaviors of systems in the smart building domain. The obtained global model from different views depicted correlations between different views.

MST-MVS proposed in paper V uses a different integration approach to obtain the global model. Cluster representatives from all the views are used in this process. 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. For each representative qualified to be used, its attribute values from other views are also extracted. An integrated matrix is built using this information (representatives of different clusters from each view and their corresponding attributes from other views). This is followed by obtaining a final global model by clustering the data points in the matrix using MST based clustering algorithm. The pre-evaluation of clustering models in each view is done to ensure that the quality of the local clustering model does not negatively affect the global model. It is important to maintain the quality of the global model 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. Figure 4.3 (taken from Paper V) illustrates how knowledge is transferred from one chunk to another in *MST-MVS* clustering algorithm.

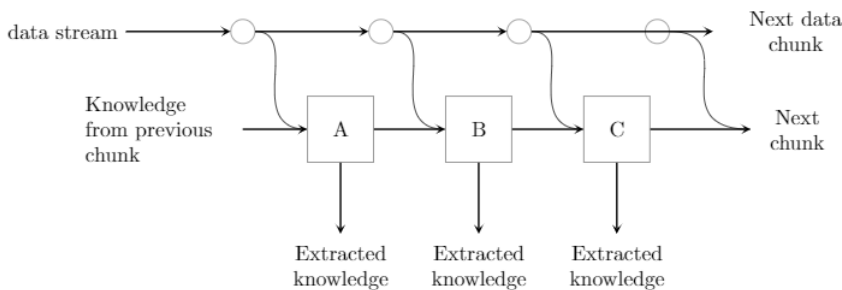


Figure 4.3: High-level overview of MST-MVS clustering algorithm for data chunks A, B, and C showcasing how knowledge from the global model of the previous chunk is used in building the clustering model of the new data chunk. Figure is taken from Paper V.

MST-MVS 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. The results indicate that the transfer of knowledge from the global model to build the local clustering solutions of the 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 to label each view's data points into different clusters obtained in the global model. This labeling technique has produced better results than the Convex Non-negative Matrix Factorization (CNMF) based labeling technique used in the study.

Paper VIII proposes a data analysis approach for high-dimensional multi-source data with missing values. A multi-layered clustering approach followed by hypergraph clustering based on k-medoids and nearest-neighbor similarity is used to achieve this. The approach is developed to be able to effectively use data with missing values, reducing information loss. This is done by using the layering approach, i.e., data objects having missing values in a layer are removed, but it can be noted that they are still being used in other layers where the features are available. It can be noted that the other multi-view clustering algorithms proposed in the study, except paper V, also work with missing and heterogeneous data, but they are not evaluated concerning this aspect in those studies. The main purpose of the global models in these scenarios is to showcase the relations between the local models; this can not be achieved for the view with missing values. Figure 4.4 (taken from Paper VIII) presents the overview of the different steps of the proposed approach. The proposed approach is tested in a real-world use case related to condition monitoring of industrial assets, i.e., a fleet of compressors used in very different conditions, like factories and hospitals, in this case. It is difficult to analyze and mine knowledge from such data as they have different technical specifications and other characteristics, thus making it difficult to compare. The approach helps to group these assets into homogeneous groups, thus enabling easier analysis. The study uses high-dimensional metadata with missing values to obtain these groups. The obtained groups are further validated using the time-series data generated during their operation.

Application Scenarios: The multi-source nature of data is common in fields like IoT where data is collected using multiple sensors, or even in other scenarios where the data is collected from different sources. In this thesis, the developed data mining techniques are evaluated on real-world use cases in the areas of smart buildings and performance monitoring of industrial assets. It can further be used in domains like health care, where data is commonly heterogeneous and available in different formats like numerical data, images, etc. Data in this area is also prone to missing values for reasons such as manual entry. The available information (feature set) can be divided into different views, different learning algorithms suitable for the data in that view can be used, and then relationships between them or consensus knowledge can be

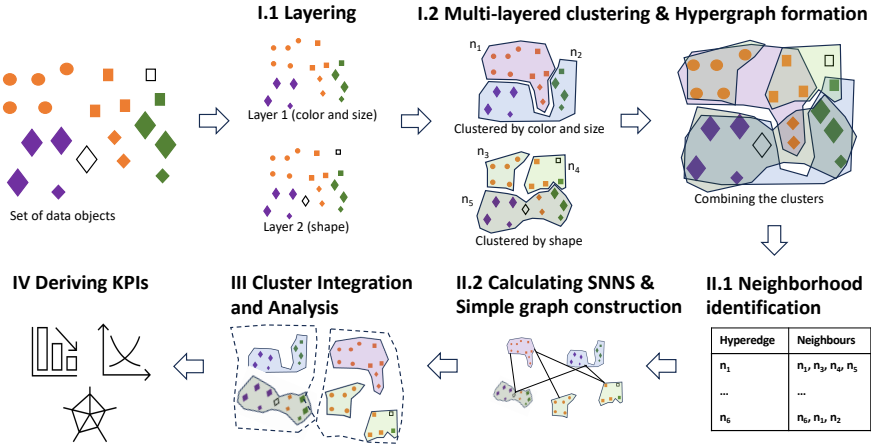


Figure 4.4: Overview of proposed data analysis approach for analyzing incomplete heterogeneous data. Figure is taken from Paper VIII.

obtained by building a global model.

4.3 Domain Adaptation

This thesis explores the area of unsupervised domain adaptation for time-series data in papers VI and VII. Novel domain adaptation algorithms based on cluster integration are proposed. Paper VI proposes *DIBCA*, which has been optimized to be robust to outliers in paper VII with *DIBCA++*. The algorithms are developed to be able to transfer new knowledge between domains. Both 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 applied on the representatives of the other domain) is used to identify the correlations. Two types of clusters are obtained from the algorithm, with common clusters presenting the behavior present in both domains and private clusters presenting the behavior of their respective domain. An integrated model that could be used across domains is obtained by combining all the common and private clusters. Along with the integrated model two personalized adapted models one for each domain consisting of common clusters and their respective private clusters, are obtained. Figure 4.5 (taken from paper VI) presents the flowchart showing different steps of *DIBCA* and *DIBCA++*.

While the clustering algorithm used by *DIBCA* is based on ISM [65] which uses the clusters' low-value and high-value vectors as representatives *DIBCA++* is developed to be robust to outliers by using the clusters' mean, standard deviation, and size. This is because *DIBCA++* makes use of all the data instances to obtain the representa-

tives unlike *DIBCA* where only minimum and maximum values of each attribute are considered to obtain the representatives. Clustering done in *DIBCA++* is inspired from [66], but unlike it where the data is assumed to be normally distributed, the proposed algorithm uses Chebyshev’s inequality which satisfies most distributions to determine the interval describing cluster boundaries. Both algorithms use only cluster representatives in their operations, thus making them resource-efficient and also preserving privacy as actual data points are not required to be disclosed.

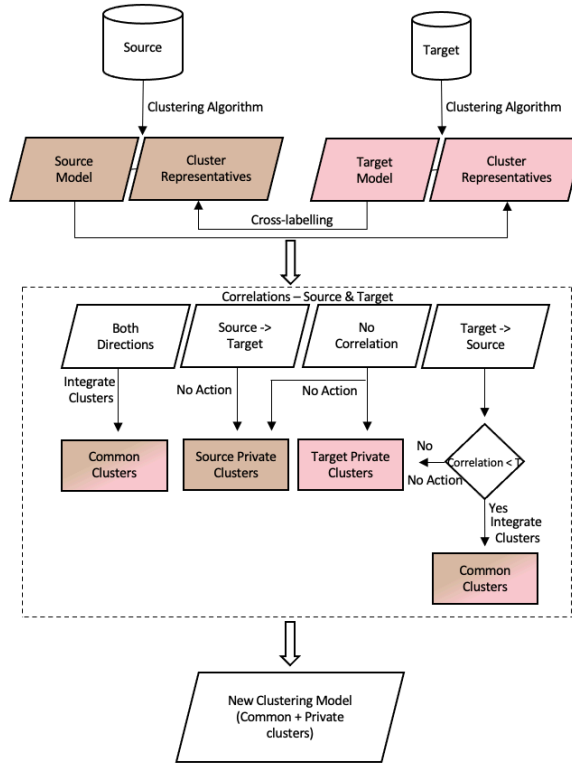


Figure 4.5: Overview of different steps of DIBCA and DIBCA++. Figure is taken from Paper VI.

DIBCA’s potential in the data labeling task is evaluated using the PAMAP2 data set, where it has performed well and correctly labeled up to 91.1% of the clusters, and in the cases where the correct match was not found they were not labeled. In the smart logistics domain based on the obtained accuracy and F-measure values the performance of the integrated and adapted models is better or comparable to the original source and target models of the considered domain. The experimental results showcase the capability of *DIBCA++* to transfer knowledge between domains, leading to improved performance. In the smart logistics use case, the adapted source and target models produced by *DIBCA++* have performed better 70% and 80% of the time, respectively. Both the algorithms were compared against each other using the

smart logistics and DaLiAc data sets. The results demonstrate the better performance of *DIBCA++*.

Application Scenarios: Domain adaptation algorithms can be used in application scenarios where there is a requirement for the model to be adapted to be used in different (new) locations or by different people/customers that have different data characteristics from those the model has been exposed during the training. One example is the case of automatic vehicles, where the actual environment they would be used in might be different from what they have been trained in. For example, if the model is trained in a location with a hot climate, but the car needs to be used in a country with a cold climate and snowfall. The model must be adapted to the new circumstances to perform well, which can be done using domain adaptation. This thesis evaluates the proposed domain adaptation algorithms in the smart logistics use case to correctly identify and perform GNSS activation as required based on the tracker's current location. This helps reduce energy consumption and localize trackers with low energy capacities. The tracker can be used in various locations like cities and countryside with many differences like the number of cell towers, traffic, etc.; the model should be able to adapt to these new circumstances. We have also tested the algorithms in the case of HAR, where knowledge obtained from training the model on one person's data could be transferred to another.

4.4 Summary

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

In Papers, I, II III, and IV 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 the newly arriving data. Techniques like bipartite graphs and MI clustering are used to achieve this. Unlike these, 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, III, and IV of this thesis, FCA integrates the clustering models from different views and builds a consensus clustering (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 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. It can be noted that only representatives of clusters that have passed the quality check are used. This approach successfully integrated information from different views and built a global model.

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

Algorithms *DIBCA* and *DIBCA++* are proposed to address this research question in papers VI and VII. These algorithms find correlations between the domains using cross-labeling (cluster representatives in one domain are labeled using the clustering model of the other domain), followed by obtaining common and private clusters based on the correlation obtained. Three different models are obtained using these clusters: one integrated model containing all the common and private clusters that could be used in both domains and two adapted (source/target) models for their respective domains. These algorithms use only cluster representatives in their operations, making them resource-efficient.

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, where a novel hybrid clustering workflow is proposed. The approach uses multi-layered clustering to obtain a hypergraph, followed by hypergraph clustering using k-medoids and shared nearest-neighbor similarity. The multi-layered clustering assists in handling missing values, reducing information loss. This clustering in each layer is done separately, and data objects with missing values in a particular layer are removed but are still used in other layers where their features are available. Other multi-view clustering algorithms (except *MST-MVS*) proposed in this thesis are also capable of handling missing values, but this feature of the algorithm is not evaluated in these studies.

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.

An evolving clustering approach, entitled *Split-Merge Evolutionary Clustering*, capable of continuously updating the generated clustering solution in the presence of new data, is proposed in Paper I. The results show the ability of the proposed algorithm to integrate newly arriving data chunks into the existing clustering model. New challenges have been studied and addressed through the progression of the work. Namely, the *Split-Merge Evolutionary 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, and VIII).

The *MV Split-Merge Clustering* (Paper II) developed can handle multi-view streaming data. The clustering model and relationships between different views obtained by the algorithm are comparable to those obtained in the batch scenario. The algorithm proposed in Paper III improves the performance and interpretability of the algorithm proposed in Paper II. 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 containing 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 Paper IV, the algorithm proposed in Paper III is evaluated in an industrial use case of smart buildings. The results show that the algorithm can be used in the smart-building domain for tasks like monitoring and analyzing. It has also successfully detected some previously known and unknown deviating behaviors of the system during the monitored period. Paper V 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. Experimental results show

that the knowledge transfer has positively impacted the generation of local clustering solutions.

Papers VI and VII propose clustering techniques for the domain adaptation problem. *DIBCA* proposed in Paper VI has performed well in automatic data labeling on a publicly available human activity recognition data set. The clustering models generated perform better or are comparable to the respective original models in the domain. *DIBCA++* has demonstrated its capability in transferring knowledge between the domains and the need/advantage for personalizing the models for each domain. Paper VIII presents a novel hybrid clustering technique for grouping heterogeneous data with many missing values to obtain homogeneous groups suitable for continuous monitoring and deviating behavior detection. Its ability is evaluated in an industrial use case of performance monitoring of assets where the algorithm was able to successfully obtain homogeneous groups with similar data characteristics.

As a part of future work, we would like to study, analyze, and test how the evolving clustering algorithms proposed/used in papers I, II, III, IV, and V perform on different types of concept drift. According to our initial analysis, all the algorithms developed in this thesis are naturally adaptive to concept drift scenarios. The algorithm automatically adapts to new data characteristics, making it difficult to detect where the concept drift occurred. We plan to explore this aspect and work in the direction of being able to detect concept drifts. We also plan to simulate different concept drift scenarios and test the algorithm's performance in these different scenarios.

Explainability is an important feature to consider while developing machine learning algorithms to be able to be accepted and used in a wide range of applications, especially if the models will be applied in fields like medicine or law. Two of the algorithms included in the current thesis specially study and showcase the explainability aspect of our approaches (Papers IV and VII). Our future interests are also directed to the area of studying explainable AI solutions. We are interested in exploring whether the proposed domain adaptation approaches could be combined with deep learning models to understand whether such combinations would improve the explainability of transfer learning or domain adaptation models based on deep learning.

6 Experiences and Learning Outcomes

This PhD 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 my PhD study, I have been involved in three different research projects and had the opportunity to work with different researchers, including practitioners from our industrial partners. I also got a chance to go on a research visit to the EluciDATA lab, Sirris in Belgium. This has exposed me to the challenge of integrating into a new research environment and has further helped me create new international collaborations. The work done as a part of this research visit has been successfully published as Paper VIII. Through my PhD, I got an opportunity to work on different interesting research problems, which are also very relevant for different application scenarios in the industry. I worked with real-world data obtained from our industrial partners in domains like smart buildings, smart logistics, and performance monitoring of industrial assets. This has provided us an opportunity to test how our proposed algorithm works on real-world data. While working with real-world data, I understood the importance of knowing the system and data well to be able to use the developed clustering models successfully. 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.

A part of the PhD was done during the COVID pandemic. The pandemic has completely changed everything and was unpredictable. Just like everyone, we were forced to adapt to new situations. Working from home, virtual meetings and conferences have become a new normal. During these tough times, I have learned to adapt to new situations and deliver. This period demanded a lot of motivation. 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 overcome my fears to a certain extent. 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] A. Zubaroglu 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.
- [5] 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.
- [6] B. Jiang and et al. “Evolutionary multi-objective optimization for multi-view clustering”. In: *2016 IEEE CEC 2016*. 2016, pp. 3308–3315.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] V. Wenz, A. Kesper, and G. Taentzer. “Clustering Heterogeneous Data Values for Data Quality Analysis”. In: *J. Data and Information Quality* 15.3 (2023).
- [12] L. Fu, P. Lin, A. V. Vasilakos, and S. Wang. “An overview of recent multi-view clustering”. In: *Neurocomputing* 402 (2020), pp. 148–161.

- [13] D. Gamberger et al. “Multilayer Clustering: A Discovery Experiment on Country Level Trading Data”. In: *Discovery Science*. Springer Int. Publ., 2014, pp. 87–98.
- [14] G. Pio, F. Serafino, D. Malerba, and M. Ceci. “Multi-type clustering and classification from heterogeneous networks”. In: *Information Sciences* 425 (2018), pp. 107–126.
- [15] M. C. de Goeij, M. van Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker. “Multiple imputation: dealing with missing data”. In: *Nephrology Dialysis Transplantation* 28.10 (2013), pp. 2415–2420.
- [16] A. Bouchachia. “Evolving clustering: an asset for evolving systems”. In: *IEEE SMC Newsletters* 36 (2011).
- [17] E. Lughofer. “A dynamic split-and-merge approach for evolving cluster models”. In: *Evolving Systems* 3.3 (Sept. 2012), pp. 135–151.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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>.
- [22] 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.
- [23] 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: *IJ-CAI*. Vol. 7. 2007, pp. 2689–2694.
- [24] 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.

- [25] 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>.
- [26] 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.
- [27] 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.
- [28] Y. Ye and et al. “Incomplete Multiview Clustering via Late Fusion”. In: *Computational Intelligence and Neuroscience* 2018 (Oct. 2018), pp. 1–11.
- [29] S. Wang and et al. “Multi-view Clustering via Late Fusion Alignment Maximization”. In: *Proceedings of IJCAI-19*. July 2019, pp. 3778–3784.
- [30] 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.
- [31] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. “A Comprehensive Survey on Transfer Learning”. In: *Proceedings of the IEEE* 109.1 (2021), pp. 43–76. DOI: 10.1109/JPROC.2020.3004555.
- [32] M. AlShehhi, E. Damiani, and D. Wang. “Toward Domain Adaptation for small data sets”. In: *Internet of Things* 16 (2021), p. 100458. ISSN: 2542-6605. DOI: <https://doi.org/10.1016/j.iot.2021.100458>.
- [33] G. Csurka. *Domain adaptation in computer vision applications*. Cham: Springer International Publishing, 2017.
- [34] S. Tang, Y. Zou, Z. Song, J. Lyu, L. Chen, M. Ye, S. Zhong, and J. Zhang. “Semantic consistency learning on manifold for source data-free unsupervised domain adaptation”. In: *Neural Networks* 152 (2022), pp. 467–478. ISSN: 0893-6080.
- [35] M. Zhu. “Source Free Domain Adaptation by Deep Embedding Clustering”. In: *2021 18th Int. Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. 2021, pp. 309–312. DOI: 10.1109/ICCWAMTIP53232.2021.9674068.
- [36] Y. Madadi, V. Seydi, K. Nasrollahi, R. Hossieni, and T. Moeslund. “Deep Visual Unsupervised Domain Adaptation for Classification Tasks: A Survey”. In: *IET Image Processing* 14.14 (2020), pp. 3283–3299. ISSN: 1751-9659. DOI: 10.1049/iet-ipr.2020.0087.

- [37] S. Hundschell, M. Weber, and P. Mandl. “An Empirical Study of Adversarial Domain Adaptation on Time Series Data”. In: *Artificial Intelligence and Soft Computing*. Ed. by L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J. M. Zurada. Cham: Springer International Publishing, 2023, pp. 39–50. ISBN: 978-3-031-23492-7.
- [38] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. “Federated Learning: Challenges, Methods, and Future Directions”. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60. DOI: 10.1109/MSP.2020.2975749.
- [39] 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.
- [40] G. W. Flake, R. E. Tarjan, and K. Tsioutsoulis. “Graph clustering and minimum cut trees”. In: *Internet Mathematics* 1.4 (2004), pp. 385–408.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] M. Zhang and Z. Zhou. “Multi-instance clustering with applications to multi-instance prediction”. In: *Applied Intelligence* 31 (2009), pp. 47–68.
- [45] B. Ganter, G. Stumme, and R. Wille. “Formal Concept Analysis: Foundations and Applications”. In: LNAI, no. 3626, Springer-Verlag, 2005.
- [46] 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.
- [47] 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.
- [48] 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.

- [49] J. A. Blackard, D. J. Dean, and C. W. Anderson. *UCI Machine Learning Repository*. 1998. URL: <http://archive.ics.uci.edu/ml>.
- [50] 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.
- [51] 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.
- [52] 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).
- [53] P. Fränti and S. Sieranoja. *K-means properties on six clustering benchmark datasets*. 2018. URL: <http://cs.uef.fi/sipu/datasets/>.
- [54] A. Reiss and D. Stricker. *Introducing a New Benchmarked Dataset for Activity Monitoring*. 2012. URL: <https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring>.
- [55] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier. “Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset”. In: *PloS one* 8.10 (2013), e75196. DOI: 10.1371/journal.pone.0075196.
- [56] K. A. Jain and C. R. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.
- [57] J. Handl, J. Knowles, and D. Kell. “Computational cluster validation in post-genomic data analysis”. In: *Bioinformatics* 21.15 (2005), pp. 3201–3212.
- [58] H. Van der Hoef and M. J. Warrens. “Understanding information theoretic measures for comparing clusterings”. In: *Behaviormetrika* 46 (2019), pp. 353–370.
- [59] “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.
- [60] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, 2002.

- [61] 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.
- [62] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. “Planning”. In: *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 89–116. ISBN: 978-3-642-29044-2. DOI: 10.1007/978-3-642-29044-2_8. URL: https://doi.org/10.1007/978-3-642-29044-2_8.
- [63] 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.
- [64] 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.
- [65] D. L. Iverson. “Inductive System Health Monitoring.” In: *IC-AI*. 2004, pp. 605–611.
- [66] P. Davidsson. “Coin Classification Using a Novel Technique for Learning Characteristic Decision Trees by Controlling the Degree of Generalization”. In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. 1996.