GRAPH-BASED ALGORITHM WITH SELF-WEIGHTED AND ADAPTIVE NEIGHBOURS LEARNING FOR MULTI-VIEW CLUSTERING

HE YANFANG

UNIVERSITI SAINS MALAYSIA

2024

GRAPH-BASED ALGORITHM WITH SELF-WEIGHTED AND ADAPTIVE NEIGHBOURS LEARNING FOR MULTI-VIEW CLUSTERING

by

HE YANFANG

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

November 2024

ACKNOWLEDGEMENT

I am deeply indebted to my supervisor, Associate Professor DR. Nurul Hashimah Binti Ahamed Hassain Malim, and Associate Professor Dr. Umi Kalsom Yusof, thank you for your unconditional support, encouragement, and specific time, energy, and opinion that have been given throughout my study in accomplishing this proposal thesis. I thank them for showing me how to identify interesting problems and how the research can start and finish correctly.

I would also like to express my appreciation to my beloved husband, Liang Shutian, who has encouraged and supported me throughout this journey. I also would like to thank my research group teams for their help and moral support that can help me to finish this proposal writing. I greatly value their friendship and sincerely appreciate their concern for me.

Above all, none of this would have been possible without the love and patience of my family, who have been a constant source of love, concern, support, and strength for all these years. Last but not least, I thank my parents for their undivided support. Without their permission, I would not get confident and motivated to finish my research. Thank you for all.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF SYMBOLS

m	number of view
k	the number of neighbours
U	Unified graph matrix
S	Initial graph matrix
c	the number of clusters
F	Initial graph matrix
$\|\| \bullet \|\|_1$	L1-norm
$\|\| \bullet \|\|_2$	L2-norm
$\|\| \bullet \|\|_F$	Frobenius
Tr	Trace
X	Dataset
w	weight
λ	initial parameters
t	anchor points number
Z	Bipartite Graph

LIST OF ABBREVIATIONS

MVC	Multi-View Clustering
MKL	Multi-Kernel learning
SDP	Semidefinite Programming
DiMSC	Diversity-induced Multi-view Subspace Clustering
LT-MSC	Low-rank Tensor Constrained Multi-view Subspace Clustering
SSC	Sparse Subspace Clustering
LRSC	Low-Rank Subspace Clustering
LRSSC	Low-Rank Sparse Subspace Clustering
DCSC	Discriminative and Coherent Subspace Clustering
CAGrad	Conflict Average Gradient descent
CGD	Cross-view Graph Diffusion
CLR	Constrained Laplacian Rank
AMCSE	Automatic Weighted Multi-view Clustering
KNN	K-Nearest Neighbor
SWCAN	Self-Weighted Clustering with Adaptive Neighbours
SG	Similarity Graph
LPP	Locality Preserving Projections
KKT	Karush Kuhn Tucker
SM	Similarity Matrix
SG	Similarity Graph
MMV	multiple measurement vector
SOCP	Second-Order Cone Programming

SDP	Semi-Definite Programming
SwMC	Parameter-free Self-weighted Multi-view Projection Clustering
AMGL	Parameter-Free Auto-Weighted Multiple Graph Learning
MVKKM	Multi-View Kernel K-Means
SWMCAN	Self-Weighted graph-based Multi-view Clustering with Adaptive Neighbours learning
SWMCAN-JG	SWMCAN model using Joint Graph learning
SWMCAN-JGBG	SWMCAN model using Bipartite Graph for large-scale data
Ncut	Normalized cut
SMSC	Self-taught Multi-view Spectral Clustering
AUMFS	Adaptive unsupervised multi-view feature selection
MVGL	graph learning for multi-view clustering
OPMC	One-pass Multi-view Clustering for Large-scale Data
BMVC	Binary multi-view clustering
LDA	Linear Discriminant Analysis
SP	Sparse Representation
MVSpec	Multi-view spectral clustering
PwMC	Parameter-weighted Multi-view Clustering
SwMC	Self-weighted multi-view clustering
GBS	general Graph-Based System
GMC	Graph-based multi-view clustering
FgMVC	Fine-grained multi-view clustering with robust multi-prototypes representation
MVASM	Multi-View clustering with Adaptive Sparse

LMBSG	Large-Scale Multi-View Spectral Clustering via Bipartite Graph
LMVSC	Large-Scale Multi-View Subspace Clustering in Linear Time
EOMSC-CA	Efficient One-Pass Multi-View Subspace Clustering with Consensus Anchors
UDBGL	Efficient Multi-view Clustering via Unified and Discrete Bipartite Graph Learning
NMI	Normalized Mutual Information
ACC	Accuracy
ARI	Adjusted Rand Index

LIST OF APPENDICES

ALGORITMA BERASASKAN GRAF DENGAN PEMBELAJARAN BERWAJARAN SENDIRI DAN JIRAN BOLEH SUAI UNTUK PENGGUGUSAN PELBAGAI PANDANGAN

ABSTRAK

Perkembangan pesat dalam teknologi perkakasan telah menghasilkan jumlah data pelbagai pandangan yang besar dengan format perwakilan yang berbeza. Walau bagaimanapun, data pelbagai pandangan yang dikumpul dari aplikasi praktikal sering terjejas daripada hingar disebabkan oleh pelbagai faktor dalam persekitaran semulajadi, menyebabkan kecabaran untuk mendapat set data yang berkualiti tinggi. Untuk menangani masalah hingar dalam data multi-pandangan, kajian ini mempertingkatkan kaedah GBS dan mengembangkan algoritma pengelompokan multi-pandangan graf berwajaran sendiri (SWMCAN) baharu. Khususnya, SWMCAN menangani hingar dalam data multi-pandangan melalui norma L1-dan mengoptimumkan fungsi objektif melalui kaedah penimbangan semula berulang yang baru. Eksperimen yang meluas dibuat pada set data sintetik dan dunia sebenar menunjukkan bahawa algoritma SWMCAN mengatasi prestasi kaedah pengelompokan multi-pandangan yang baru dicadangkan dari segi prestasi pengelompokan dan ketahanan terhadap hingar secara konsisten. Walaupun algoritma SWMCAN menyelesaikan masalah hingar dalam data berbilang paparan, graf awal dan akhirnya adalah bebas dan tidak boleh belajar antara satu sama lain. Oleh itu, kajian ini menggabungkan pembelajaran graf bersama daripada algoritma GMC ke dalam SWMCAN dan mencipta algoritma baharu yang dipanggil SWMCAN-JG untuk menangani isu ini. Algoritma SWMCAN-JG berkesan menangani kedua-dua masalah bunyi dan kebebasan secara serentak. Khususnya, SWMCAN-JG berbeza dengan GMC, ia menggunakan norma L1-sebagai matriks baru untuk mengukur jarak graf akhir dan mengoptimumkan algoritma melalui kaedah penimbangan semula berulang. Eksperimen yang melibatkan set data sintetik dan dunia sebenar juga menunjukkan SWMCAN-JG mengatasi kaedah penggugusan pelbagai pandangan terkini secara konsisten dari segi prestasi penggugusan dan kekukuhan mengatasi hingar. Walauba-

gaimanapun, algoritma SWMCAN dan variasinya, SWMCAN-JG, tidak sesuai untuk mengendalikan set data yang berskala besar kerana kerumitan pengiraan yang tinggi dan masa pelaksanaan yang panjang. Dengan ini, kajian ini menambah-baik model SWMCAN dengan menggunakan Graf Bipartit (SWMCAN-JGBG) supaya menyelesaikan kerumitan pengiraan yang tinggi dalam penggugusan pelbagai pandangan apabila melibatkan data berskala besar. Algoritma ini menggabungkan rangka kerja LMSBG dengan proses penjanaan matriks persamaan daripada algoritma SWMCAN-JG untuk membentuk Graf Bipartit yang direka khusus untuk set data yang besar. Pendekatan ini menggunakan satu set wakil jangkar bersatu yang ringkas untuk pelbagai pandangan untuk mengenkapsulasi maklumat konsensus. Sebuah graf bipartit dibentuk yang menghubungkan titik data dan jangkar ini. Eksperimen yang melibatkan set data dunia sebenar menunjukkan bahawa pendekatan ini mempunyai keberkesanan dan kecekapan yang lebih unggul berbanding dengan kaedah lanjutan lain apabila memproses set data pelbagai pandangan yang berskala besar. Kaedah SWMCAN-JGBG yang dicadangkan kami berjalan sehingga hampir 1116 kali lebih pantas daripada kaedah EOMSC-CA tercanggih.

GRAPH-BASED ALGORITHM WITH SELF-WEIGHTED AND ADAPTIVE NEIGHBOURS LEARNING FOR MULTI-VIEW CLUSTERING

ABSTRACT

The rapid advancement in hardware technology has generated a substantial volume of multi-view data with diverse representation formats. However, in practical applications, the collected multi-view data is often affected by noise due to various factors in the natural environment, making it challenging to obtain a high-quality dataset. To address the noise problem in multi-view data, this study enhances the GBS method and develops a new self-weighted graph multi-view clustering algorithm (SWMCAN). Particularly, SWMCAN addresses multi-view data noise using the L1-norm and optimizes the objective function through a novel iterative reweighted method. Extensive experiments on synthetic and real-world datasets consistently demonstrate that the SWMCAN algorithm outperforms recently proposed multi-view clustering methods regarding clustering performance and noise robustness. Although the SWMCAN algorithm solves the noise problem in multi-view data, its initial and final graphs are independent and cannot learn from each other. To address this issue, this study incorporated joint graph learning from the GMC algorithm into SWMCAN, creating a new algorithm called SWMCAN-JG. The SWMCAN-JG algorithm effectively tackles both noise and independence problems simultaneously. Specifically, unlike GMC, SWMCAN-JG employs the L1-norm as a new metric for measuring the final graph distance and optimizes the algorithm through an iterative reweighted method. Extensive experiments on synthetic and real-world datasets consistently demonstrate that SWMCAN-JG outperforms recently proposed multi-view clustering methods regarding clustering performance and noise robustness. The SWMCAN algorithm and its variant, SWMCAN-JG, are unsuitable for handling large-scale datasets due to their high computational complexity and prolonged execution times. To tackle the high computational complexity in multi-view clustering, a new algorithm called SWMCAN-JGBG was developed. This algorithm combines the LMSBG framework with the similarity matrix generation process from the SWMCAN-JG algorithm, forming a Bipartite Graph specifically designed for large

datasets. The approach leverages a concise set of representative unified anchors for diverse views to encapsulate consensus information. A bipartite graph is established connecting data points and these anchors. Extensive experiments on real-world datasets demonstrate that, compared to other advanced methods, this approach exhibits superior effectiveness and efficiency, when processing large-scale multi-view datasets. Our proposed SWMCAN-JGBG method runs up to nearly 1116 times faster than the state-of-the-art EOMSC-CA methods.

CHAPTER 1

INTRODUCTION

1.1 Background

Single-view data refers to information from a single-view, mode, or feature space. This data type only describes an object or phenomenon, reflecting a single attribute R. Xu et al., 2005. The analysis and processing of single-view data are relatively simple due to their origin from a single source or measurement method. For example, images captured from a single camera, text data from a single source such as a book or a website, and single-type medical examination data, such as blood test results. Universal datasets typically represent objects from a single view and cannot comprehensively represent information. However, multimedia technology simplifies data collection, diversifying the sources and features of datasets. This type of data is called multi-view data.

Specifically, multi-view data are collected from different sources and fields because of the rapid development of information technology. Therefore, these data are usually presented in various forms. For example, in web mining, a web page consists of text, images, and hyperlinks; in document mining, text can be expressed in various languages, such as Korean, Japanese, English, and Chinese, which can be considered four different views of the same news M.-S. Chen et al., 2020; S. Huang et al., 2019; Mei et al., 2022, such as Figure 1.1. Each view in multi-view data has particular attributes related to knowledge discovery tasks. Different views typically contain supplementary information. Multi-view data can describe the same object from different views to represent the data comprehensively. Multi-view clustering learning can integrate all these view information simultaneously and use the complementarity and consistency between views to obtain more accurate clustering (Hou et al., 2017).

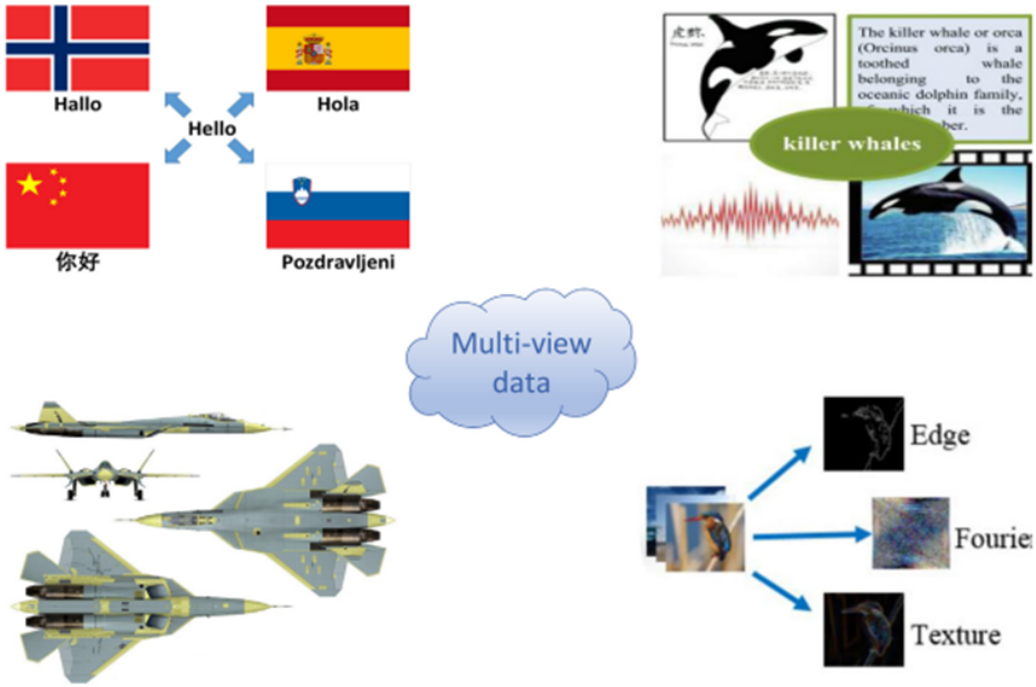


Figure 1.1: Multi-view data

Although multi-view data provides richer information and higher accuracy, it also introduces challenges, including multi-view noise and high computational complexity. During the data collection process, noise is introduced due to limitations and differences in collection equipment, changes in environmental conditions, and inherent noise in the data itself. This noise affects the data quality and interferes with subsequent data analysis and ing processes, resulting in inaccurate results. High computational complexity refers to the processing of multi-view data, which involves data fusion and analysis from multiple views, resulting in increased computational demands and longer algorithm running times. Especially for large-scale datasets, improving algorithm efficiency and reducing running time while ensuring accuracy is an urgent problem. In this study, high computational complexity is measured by running time.

For example, in medical research, multi-view data of patients (such as MRI, CT scans, and X-ray images) are collected simultaneously for diagnosis and treatment. The MRI view mainly provides high-resolution soft tissue images, the CT view offers cross-sectional images of the body's interior, and the X-ray view primarily shows images of bones and other dense tissues. Patient movement during scanning can cause image

blurring or artifacts, leading to noise in MRI images. In CT scans, improper equipment calibration or the presence of metal implants may cause artifacts or stripes, resulting in noise. Similarly, inappropriate exposure time or intensity in X-ray imaging can lead to overexposure or underexposure, causing detail loss and noise.

In hospitals, patients typically undergo various imaging examinations such as MRI, CT, X-rays, and ultrasound. These imaging data must be integrated and analyzed to provide comprehensive diagnostic information. Extracting information from multi-view images is essential for disease diagnosis, treatment planning, and postoperative evaluation. The registration of different modal images requires complex optimization algorithms to ensure alignment. Integrating image data from different modalities enhances features and improves diagnostic accuracy. However, processing multi-view large-scale data is complex, leading to long computation times and difficulties in real-time diagnosis. When processing large-scale data, algorithms with high computational complexity significantly increase running time. High computational complexity may also reduce accuracy, affecting the reliability of diagnostic results.

Clustering is a prominent focus of machine learning research (Berkhin, 2006) because it can effectively distinguish between similar and different data objects. Additionally, clustering aids in analyzing the internal structure of data and discovering patterns and relationships within datasets. Clustering is widely used in data mining (Tan et al., 2013), computer vision (Xia et al., 2016), and pattern recognition (Namratha & Prajwala, 2012). Common clustering methods, such as k-means and spectral clustering, have been crucial in practical applications (Berkhin, 2006). However, these algorithms belong to traditional single-view clustering methods, which cannot handle data containing multiple features simultaneously. Traditional clustering algorithms typically utilize a set of features or a topic information window and can only handle single-view data. Clustering techniques can be applied to multi-view data, resulting in what are known as multi-view clustering algorithms.

In recent decades, there has been increased attention on multi-view algorithms in data mining, computer vision, and bioinformatics (Fu et al., 2020). These multi-view clustering algorithms can handle multi-view data. Different views are represented by different sources in multi-view data (Fu et al., 2020), which often complement each other. Proper utilization of information from multiple views generally leads to improved cluster performance instead of solely relying on individual views for multi-view clustering. Consequently, multi-view learning in clustering scenarios has led to numerous novel multi-view clustering algorithms tailored to multi-view data. Multi-view clustering includes graph-based multi-view clustering (Mei et al., 2022), multi-kernel learning (Zhu et al., 2018), co-training (Zhu & Goldberg, 2022), multi-task learning (Y. Zhang & Yang, 2018) and multi-view subspace clustering (Gao et al., 2015). One of promising state of art methods is graph-based multi-view clustering (Z. Li et al., 2021) algorithm because it can capture significant similarity and correlation between data points to achieve more accurate than other multi-view clustering. To address the issues of multi-view noise and high computational complexity, graph-based multi-view clustering methods are gaining attention.

Graph theory offers unique advantages in handling complex relationships and data structures. Constructing graph algorithms can effectively represent the relationships and structures within multi-view data. However, traditional graph clustering methods still need to improve when dealing with multi-view noise and high computational complexity. Graph-based learning offers an efficient algorithm for ing data in clustering tasks (H. Xu et al., 2020). Conventional graph learning techniques primarily revolve around constructing graphs based on single-view data. Practical scenarios often require constructing multiple graphs for various applications. Therefore, multi-view graph learning is applied in machine learning and pattern recognition (Jia et al., 2021; Liang et al., 2022).

Many graph-based learning clustering methods have been proposed recently. These methods seek to learn graphs from data dynamically. For example, (F. Nie et al., 2014)

constructed a graph based on adaptive neighbors, where the probability of one data point being adjacent to another is used to measure similarity. Later, many researchers extended this idea to handle multi-view data. Graph-based multi-view clustering requires assigning weights to each graph (T. Wang et al., 2021). Traditionally, weight parameters are used to control the weight range, typically from 1 to 100, which is relatively large. To address this issue, (Nie et al., 2016) proposed a new self-weighted multiple graph learning (AMGL) framework that automatically learns a set of weights for all graphs without any parameters. Most graph-based clustering methods require additional clustering techniques to obtain the final results. To address this issue, (F. Nie et al., 2014) proposed a new method that imposes rank constraints on the Laplacian matrix of the learned similarity matrix. This ensures ideal neighbour allocation, where the connected components in the data accurately correspond to the number of clusters, and each connected component is correct. This new algorithm simultaneously learns the data similarity matrix and clustering structure to achieve optimal clustering results.

On the other hand, the graph-based method obtains consensus information by learning a few representative unified anchor points from different views. Each anchor point is the centroid of the corresponding sub-cluster. Each view has an anchor set, and these anchors in different views store information in the same sub-cluster. A bipartite graph is then constructed between the original data points and these anchor points. These generated points play an important role in capturing the manifold of the original view.

Although graph-based methods have introduced self-weighed, adaptive nearest neighbour learning, and rank constraints to address parameter issues, fixed similarity graph problems, and the need for additional clustering methods in multi-view clustering, they still cannot solve the noise and high computational complexity problems in multi-view clustering.

1.2 Motivation

Machine learning and data mining technologies are mainly used to process data essential for the daily lives and industrial production. Therefore, data is an important cornerstone in machine learning and artificial intelligence (Winston, 1992). These technologies enable the discovery of underlying patterns and relationships within the data, thereby facilitating the extraction of valuable knowledge. Practical, novel, potentially useful, and ultimately understandable knowledge can enhance our lives and production, enabling the realization of equipment automation, system intelligence, and ecosystem intelligence.

Extensive research has historically focused on numerous advanced clustering algorithms. While these clustering algorithms have succeeded considerably, most are designed solely for single-view data (Y. Yang & Wang, 2018). Traditional clustering algorithms typically focus on homogeneous data types, describing their characteristics through a single set of attributes or relationships. In recent years, advancements in internet technology have led to the emergence of datasets with more comprehensive information. These datasets encompass not only multiple data objects from diverse sources but also intricate relationships between these objects. Multi-view learning clustering technology integrates data from different views to provide more comprehensive and accurate clustering results than single-view methods. Therefore, many people study unsupervised multi-view learning (Y. Dai et al., 2019).

Multi-view clustering algorithms are currently a highly researched topic (L. Li & He, 2020; Z. Li et al., 2022; F. Nie et al., 2016). Z. Ren et al. (2021) integrated clean embedding space learning and consistent affinity graph learning into a unified objective function to address this collaborative relationship issue. The multi-view clustering method also adopts a two-step approach. The first step involves learning a fixed similarity matrix, while the second step focuses on generating the final matrix using the learned similarity matrix. To address the two-step problem, (Tang et al., 2022)

proposed a one-step multi-view spectral clustering method, which combines k-means and spectral embedding within a unified framework to derive discrete cluster labels. The authors in (Liang et al., 2019) proposed a novel graph-based multiview clustering method, which combines the fusion of similar and distance graphs (non-similar graphs) to obtain clustering results in one step. The authors in (Y.-M. Xu et al., 2016) introduced a weighted multi-view clustering and feature selection strategy emphasizing different views' influence on clustering. This strategy assigns weights to the data point views and feature representations within each view. The authors in (Tzortzis & Likas, 2012) presented an algorithm that assigns weights to each kernel matrix, where each matrix representation corresponds to a specific view. However, these algorithms require manual adjustment of parameters to control the weights of different views. In order to solve this parameter problem, existing multiview clustering algorithms automatically assign weights for each view. For example, the authors in (S. Huang et al., 2019, 2020; Shi et al., 2020) proposed algorithms that accomplish the multi-view clustering task and learn the similarity relationship in the kernel space. Another challenge most graph-based multi-view clustering algorithms rely on additional k-means algorithms to obtain the final clustering results from the unified matrix. To achieve clustering results directly, the authors in (Qiang et al., 2021) proposed a rapid multi-view discrete clustering method incorporating anchors, efficiently solving the spectral clustering problem that can automatically obtain aggregation graphs.

Current multi-view clustering algorithms have addressed some issues, but significant challenges remain in handling multi-view data noise and computational complexity. The noise in multi-view data significantly impacts data quality and clustering accuracy. Existing methods often perform poorly in handling noisy data, affecting the robustness and stability of clustering results. Additionally, the high computational complexity of traditional clustering algorithms has become a major bottleneck with the continuous expansion of data scale, limiting their application to large-scale datasets. Therefore, developing novel multi-view clustering methods to effectively handle noisy data and reduce computational complexity is urgently needed. This will improve clustering

quality and expand the application scope of these algorithms.

This study aims to address the shortcomings of current multi-view clustering methods in handling multi-view noise and high computational complexity. We aim to better deal with multi-view data noise by improving existing algorithms, thereby enhancing clustering accuracy and robustness. The algorithm improvements will also address high computational complexity, enabling efficient operation on large-scale datasets and reducing running time.

To address the problem of multi-view data noise, this study improves the existing objective function and introduces self-weighted, rank constraint, and L1-norm. Self-weighted technology adaptively allocates weights to data from different views based on their importance. Rank constraints directly generate clustering results without additional clustering methods, and the L1-norm enhances the algorithm's sparsity, effectively suppressing the influence of noise. To address the issue of high computational complexity, this study improves the existing objective function, introduces self-weighted techniques, and employs bipartite graphs to reduce further computational complexity and running time.

1.3 Problem Statement

In multi-view data analysis, data from different views provide rich information that enhances the accuracy and comprehensiveness of clustering results. However, multi-view data often contain noise, which can significantly affect data quality and reduce the accuracy and robustness of clustering outcomes. Most existing multi-view clustering methods treat noisy data equally with normal data, ignoring the unique properties of noise and outlier data. Consequently, existing multi-view clustering is often sensitive to noisy data. The previous studies agreed that the performances of almost all the state-of-the-art multi-view clustering algorithms degraded significantly due to noisy data. Zong et al. (2018) presented a novel weighted multi-view spectral clustering algorithm,

employing spectral perturbations to algorithm the weights of different views. J. Liu et al. (2020) proposed a novel algorithm, the clustering weighted kernel k-means, for multi-view clustering. This approach assigns weights to internal clusters in each view by considering their intra-cluster similarity with corresponding clusters in other views, giving higher weights to clusters demonstrating more pronounced intra-cluster similarity. Typically, these algorithms (J. Liu et al., 2020) involve learning weights by introducing additive hyperparameters (weight-parameter), necessitating an extensive search for the optimal value. Although these multi-view clustering methods improve clustering accuracy by assigning weights to each view and fusing all views, their accuracy is low when dealing with noisy multi-view data.

The most advanced multi-view clustering technology currently available is graph-based clustering. Recently, (H. Wang, Yang, Liu, & Fujita, 2019) proposed a general graph-based multi-view clustering system (GBS). GBS's working principle involves extracting each view's data feature matrix, constructing the graph matrix for all views, and fusing these matrices to generate a unified graph matrix, thereby obtaining the final clustering. A multi-view clustering method suitable for the GBS framework has been proposed, which can (1) Effectively construct a data graph matrix, (2) Automatically weight each graph matrix, and (3) Directly generate clustering results. Experimental results on benchmark datasets show that this method significantly outperforms existing baselines. Although this method is a relatively advanced approach with high clustering accuracy, its performance is poor when dealing with multi-view data containing noise.

Our study builds on GBS, retaining its strong performance while addressing the issue of multi-view data noise. We improve GBS by modifying its objective function. This GBS objective function $\min_{S} \sum_{v=1}^{m} w_v ||S-A||_2$, used to measure the difference between two graphs, S and A, assesses the numerical distance between them. Specifically, matrices S and A represent the similarity matrix and affinity matrix of two graphs, respectively. In many algorithms, this difference metric is incorporated into the objective function to minimize the difference by adjusting S, thereby making the fused graph as close as

possible to the initial graph or meeting certain optimization conditions. To address multi-view data noise in the GBS objective function, we modify this difference metric by replacing the L2-norm with the L1-norm. B. Cheng et al. (2010) proposed the L1-norm graph algorithm to solve the noise problem. The L1-norm graph algorithm was used to solve the adaptive neighbourhood issue, and the graph is data noise-resistant. Nie et al. (2016) proposed the constrained Laplacian rank algorithm to learn a graph using the L1-norm and L2-norm. However, these L1-norm and L2-norm construct a graph structure under a single-view. While the L1-norm has been explored for mitigating noise issues in single-view clustering, its application to address noise problems in multi-view clustering remains unexplored. Replacing the L2-norm with the L1- norm effectively addresses the noise problem in multi-view data analysis. Due to the L1-norm's low sensitivity to outliers, the algorithm performs more stably and accurately when handling noisy data.

On the other hand, GBS uses the semi-definite programming(SDP) method to solve the objective function. Existing algorithms typically reformulate it as a second-order cone programming (SOCP) or semi-definite programming (SDP) problem, which can be solved using the interior point or bundle method. However, solving SOCP or SDP is computationally expensive, limiting their practical use. Recently, a method was proposed to address specific problems. This efficient algorithm rephrases the problem as a min-max problem and then applies the proximal method. Experimental results show that this algorithm is more efficient than existing algorithms. However, it is a gradient descent algorithm with a slow convergence speed. We can solve the objective function based on the L1-norm minimization problem using a straightforward, easy-to-implement, efficient iterative reweighted method.

The new objective function based on the L1-norm we modified in GBS also has some limitations. Typically, these methods optimize their objectives using a fixed graph similarity matrix for all views, meaning that the corresponding graph similarity matrix is independent of the final matrix and cannot learn from each other. For example,

the similarity matrix S and the final matrix U in GBS cannot interact, decreasing clustering accuracy. The authors in (Nie et al., 2021) developed a novel fuzzy clustering algorithm that integrates member matrix learning and anchor-based similarity graph learning within a unified framework. This method effectively utilizes prior knowledge of anchors to solve noise problems and enhance clustering performance. While most existing work focuses on merging comprehensive information from multiple views to accomplish clustering tasks, these studies often overlook the collaborative relationship between fused graphs and independent views (R. Wang et al., 2021). Our research aims to address both noise and independence issues in multi-view clustering. Inspired by the reference (H. Wang, Yang, & Liu, 2019), we adopt the joint learning method of GMC and combine it with our modified GBS objective function to form a new objective. This new objective function can simultaneously solve both noise and independence issues. The GMC algorithm performs poorly in addressing multi-view noise issues because it utilizes the L2-norm to measure the distance between the final and similar graphs. GMC also uses the semi-definite programming (SDP) method to solve the objective function. However, this method is complex in solving the objective function. Additionally, we use a new iterative reweighted method for this minimization objective function, which is distinct from GMC's approach. Our new objective function employs a simpler and easier-to-implement iterative reweighted method to solve the minimization problem.

For a dataset with n data points, graph-based clustering first constructs an $n \times n$ similarity matrix and then performs eigen decomposition on the corresponding Laplacian matrix. If the feature dimension is d, the time complexity of the first step $O(n^2 d)$, while eigen-decomposition requires $O(n^3)$ (T.-L. Liu, 2017; W. Yang et al., 2023). Therefore, conventional graph-based clustering methods become unsuitable for large-scale multi-view data as data number grows. Existing multi-view clustering algorithms struggle to handle large-scale data due to their high computational complexity, resulting in prolonged running times. Instead of relying on some $n \times n$ graph, the bipartite graph-based methods typically generate a small set of m anchors from the original data, and then construct an $n \times m$ ($m \ll n$) bipartite graph to represent the data structure

(D. Huang et al., 2019), where m is the number of anchors.

When exploring the application of graph-based multi-view clustering for noise processing, it was discovered that numerous multi-view clustering algorithms exhibit high computational complexity and encounter limitations when handling large-scale multi-view data (M.-S. Chen et al., 2020). This results in longer running times for multi-view clustering algorithms. Despite their effectiveness, graph-based clustering algorithms often face high computational complexity. To tackle this challenge, several approaches have been developed to reduce the complexity of graph construction. For example, (S. Huang et al., 2020; Y. Li et al., 2015) proposed a multi-view spectral clustering technique employing bipartite graphs to alleviate computational complexity. However, these two approaches require them to maintain a fixed input graph throughout the fusion process, compromising clustering accuracy. Inspired by reference (Y. Li et al., 2015), we modified the objective function to replace the fixed input graph with an adaptive graph, forming a new objective function. This adaptive graph comes from the objective function we modified.

1.4 Research Questions

To address the problems of multi-view clustering specified above, the following research questions are presented:

i. How to improve graph-based multi-view clustering methods to handle noise in multi-view data and efficiently solve the objective function?

ii. How to further improve graph-based multi-view clustering methods, using efficient objective function solving methods to simultaneously handle the independent learning of initial and final graphs and the noise problem of multi-view data?

iii. How to enhance an algorithm to handle large-scale data when solving the high complexity problem of multi-view clustering?

1.5 Research Objectives

This research establishes the following objectives based on the problem statement and the aim of the study:

Aim: To improvise a comprehensive self-weighted graph-based multi-view clustering with adaptive neighbours learning to ensure the efficiency, robustness, and optimal clustering performance of multi-view clustering.

i. To design the self-weighted graph-based multi-view clustering with an adaptive neighbour learning (SWMCAN) algorithm and an efficient iterative reweighted method to solve the minimization objective function, effectively addressing the noise issue in multi-view data.

ii. To improvise the SWMCAN algorithm using joint graph learning(SWMCAN-JG) and use an efficient iterative reweighted method to solve the minimization objective function to solve independent learning problems and the noise problem of multi-view data.

iii. To enhance the SWMCAN-JG method using bipartite graph for large-scale data (SWMCAN-JGBG) to solve the high computational complexity problem of large-scale data.

1.6 Research Scope

This study includes applying and improving graph-based multi-view clustering algorithms in adaptive nearest neighbours, self-weighted technology, joint graph technology, spectral clustering technology, and rank constraint. All objectives will consider multi-view clustering, self-weighted, and adaptive nearest neighbors to address the issues of noise and high computational complexity in multi-view data. The L1-norm method, reweighted iterative method, and bipartite graph are my focus to integrate

multiple views, which is different from other algorithm studies and has demonstrated its superiority in performance through experiments. This study also investigates using adaptive nearest neighbours techniques based on bipartite graphs in processing high computational complexity for large-scale data.

In addition, the performance of all proposed algorithms is based on five metrics, which will be detailed in Chapter 3. The results obtained from the experiment will be compared with other advanced multi-view algorithms and analysed using a benchmarked dataset. All three proposed objective functions are improvements of the original ones. In Chapter 3, the original method will be introduced in detail.

In order to evaluate the developed algorithms, 15 standard benchmark datasets with different views were used. Many researchers have recently used these benchmark datasets for multi-view clustering. A detailed description of the benchmark dataset can be found in Chapter 3. The effectiveness of mitigating noise in multi-view data through the proposed algorithm was showcased using robustness metrics. Noise was introduced into real-world datasets for experiments focusing on Objectives 1 and 2. Additionally, comparing the results of the objective functions for Objectives 1 and 2 in the experiment verified that Objective 2 addressed the issue of learning independence between the initial graph and the fusion graph observed in Objective 1. Furthermore, the algorithm's capacity to address computational complexity was illustrated by evaluating its performance on large-scale multi-view data, measuring runtime efficiency in Objective 3.

1.7 Thesis Outline

The flow of the thesis in chapters is outlined below: The remainder of this thesis is shown in this section. In Chapter 2, a detailed review of the relevant work can be obtained. The main contributions of the thesis are presented in Chapters 4-6. From Figure 1.2, it can be observed that Chapter 4 discusses the issue of Objective 1. Chapter

5 addresses the issue of Objective 2, while Chapter 6 addresses Objective 3. Chapter 7 provides a summary of the entire thesis.

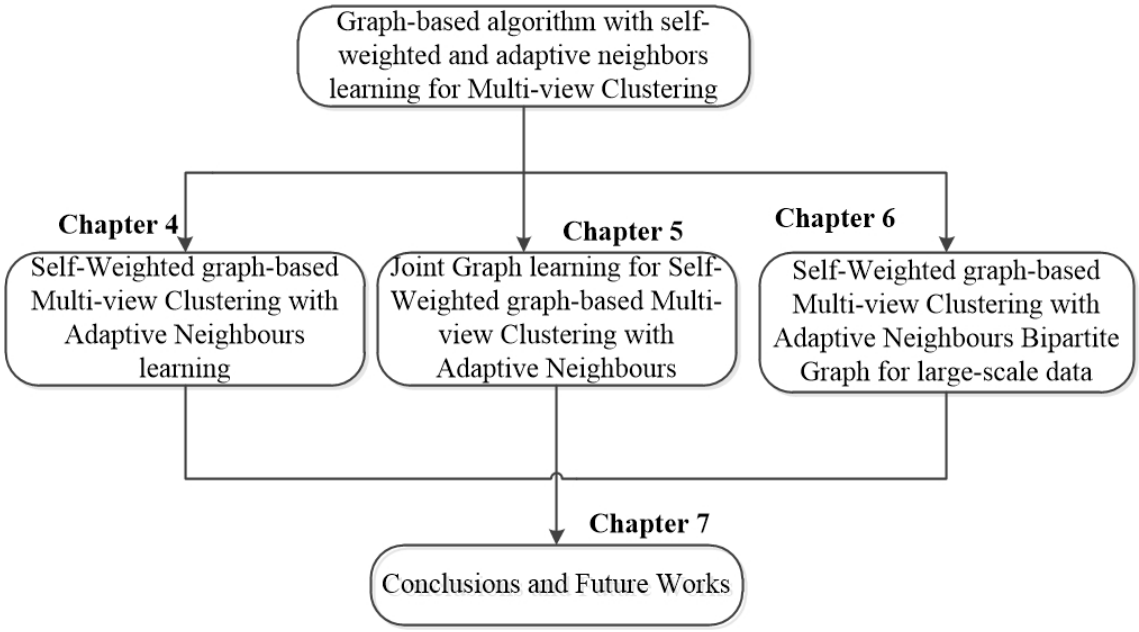


Figure 1.2: Thesis outline

i. Chapter 2: This study categorizes multi-view clustering algorithms into five classes: co-training learning, multi-kernel learning, graph-based multi-view clustering, multi-view subspace clustering, and multi-task learning. It then provides an overview of relevant graph-based clustering algorithms, encompassing rank constraints, L1-norm, L2-norm, Frobenius-norm, Joint L2,1-norm, and Adaptive Graph Learning. Subsequently, spectral clustering, the self-weighted algorithm, rank-constrained algorithms, adaptive graph learning, and algorithms designed for large-scale data in graph-based multi-view clustering are introduced.

ii. Chapter 3: This chapter introduces the overview and algorithms of the design algorithm for this study. The problem algorithm is defined, the three proposed solutions are designed, datasets are described, and the evaluation metrics used for the results are presented.

iii. Chapter 4: This chapter presents a graph-based multi-view clustering algorithm. The stability and effectiveness of the proposed algorithm are demonstrated by comparing it with five recently proposed multi-view clustering algorithms. The experimental results are verified using benchmark seven datasets.

iv. Chapter 5: A robust joint graph algorithm is proposed, integrating the similarity and consistency matrices from multiple views. The algorithm incorporates alternating iterations for achieving convergence.

v. Chapter 6: This study presents an adaptive Neighbours bipartite graph algorithm designed to address the challenges of handling high computational complexity. The algorithm's effectiveness is evaluated by comparing it to other large-scale data algorithms.

vi. Chapter 7: Conclusion and future work summarized the research findings; contributions, limitations of the study, and suggestions for further study are also discussed.

CHAPTER 2

LITERATURE REVIEWS

2.1 Introduction

This chapter provides an overview of the research on multi-view clustering algorithms and the graph-based multi-view clustering techniques applied in this study. Section 2.2 introduces the flowchart of multi-view data. Section 2.3 introduces five commonly used multi-view clustering algorithms; Section 2.4 introduces various graph-based and graph-based multi-view clustering algorithms. Section 2.5 discusses the multi-view clustering techniques used. Section 2.6 summarizes the differences between the proposed three proposed algorithms in multi-view clustering algorithms, while Section 2.7 summarizes the content of this chapter. The system content of this chapter has been organized, as shown in Fig.2.1.

2.2 Multi-view Data

In traditional data collection, all data objects in a dataset are obtained from a single-view, utilizing the same sensor or data source. Due to the data object being observed and collected solely from a single-view (Z. Li & Snavely, 2018), the resulting data is limited to that view. This type of data is referred to as single-view data in this study, which has become more prevalent with the advent of the Internet, sensors, and storage devices. With the widespread adoption of technology, data acquisition and storage have become more convenient, resulting in new data characteristics. In the era of big data, data collection from multiple views has become common. Collected data from multiple views constitute multi-view data (R. Zhang et al., 2019).

The advancement in hardware technology has led to the generation of a substantial number of multi-view data with diverse representations in practical applications.

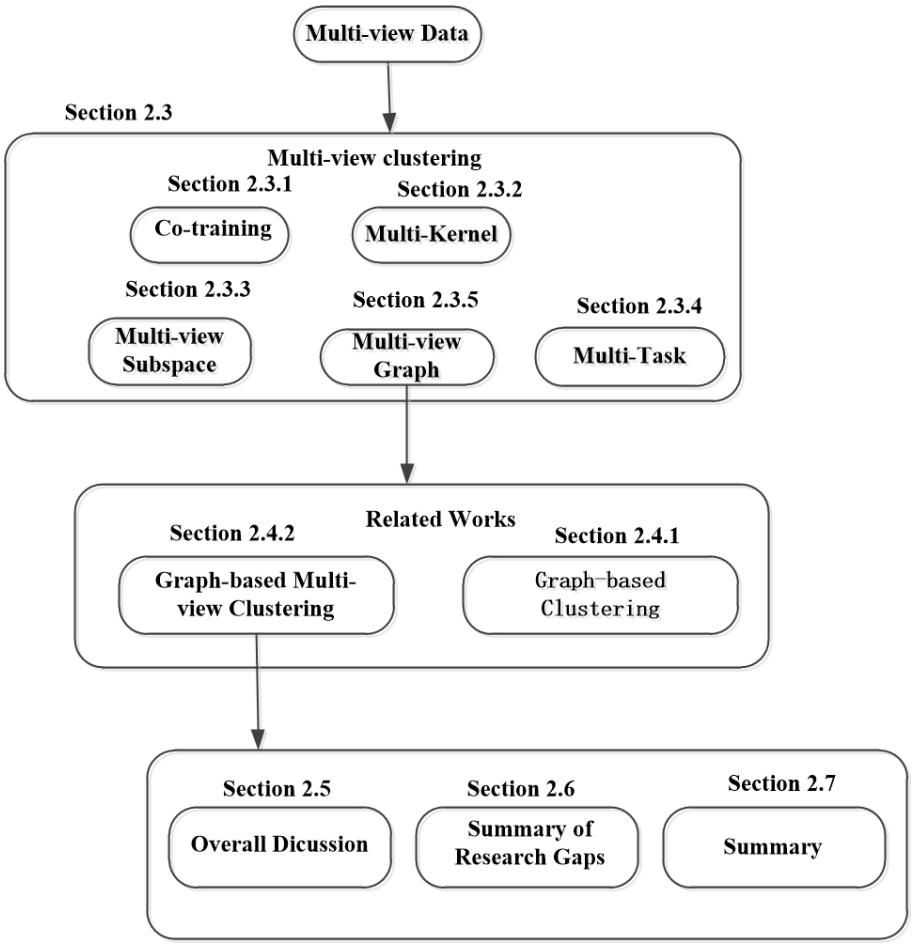


Figure 2.1: Chapter 2 content structure

Multi-view data stems from diverse sources and embodies distinct features related to various tasks. As an illustration, in image analysis, images can be described using multiple visual descriptors, including GIST, CTM, LBP, SIFT, and HOG. In web page classification, web pages can be categorized based on their content or links (S. Huang et al., 2018). Figure 2.2 provides specific examples of multi-view data.

Figure 2.2 illustrates the numerous reasons for generating multi-view data. Firstly, multi-view data can arise from the same source under varying collection conditions. For instance, in Figure 2.2a, the same camera captures the Oriental Pearl under different lighting conditions: day and night. Furthermore, the camera records the same object from various angles, the recorder captures the same person's voice under quiet and noisy conditions, and a single hydrological monitoring station measures the water level during morning and nighttime. Secondly, multi-view data can arise from different data

(a) Figure 2.2a

(b) Figure 2.2b

(c) Figure 2.2c

(d) Figure 2.2d

Figure 2.2: Detailed description of multi-view data

modes within the same source. Thirdly, multi-view data can be created from other data modes within the same source. For example, in Figure 2.2(c), a web page describing a basketball game may incorporate two modes of data: pictures and text. Additionally, multi-view data can be formed by analysing and extracting different features from the same original data. For instance, in Figure 2.2(d), a face picture can be processed to extract grayscale and LBP features, while different features like colour and texture can be extracted from the original data of the same image. Lastly, multi-view data can also be formed from various sources. For example, in Figure 2.2b, multi-view data can be obtained from data collected by different cameras on a bus and simultaneously interpreting data from various sensors. Different information systems collect user information and reports on news events generated by other media.

Retrieve multi-view data from diverse sources, including images, texts, audio, and videos. Preprocess the data from each view before employing various multi-view algorithms to amalgamate the datasets. Employ the selected clustering algorithm to algorithm and optimize the data. Multi-view clustering integrates data from various views to enhance the accuracy and robustness of clustering results. The entire process of multi-view clustering is shown in Figure 2.3:

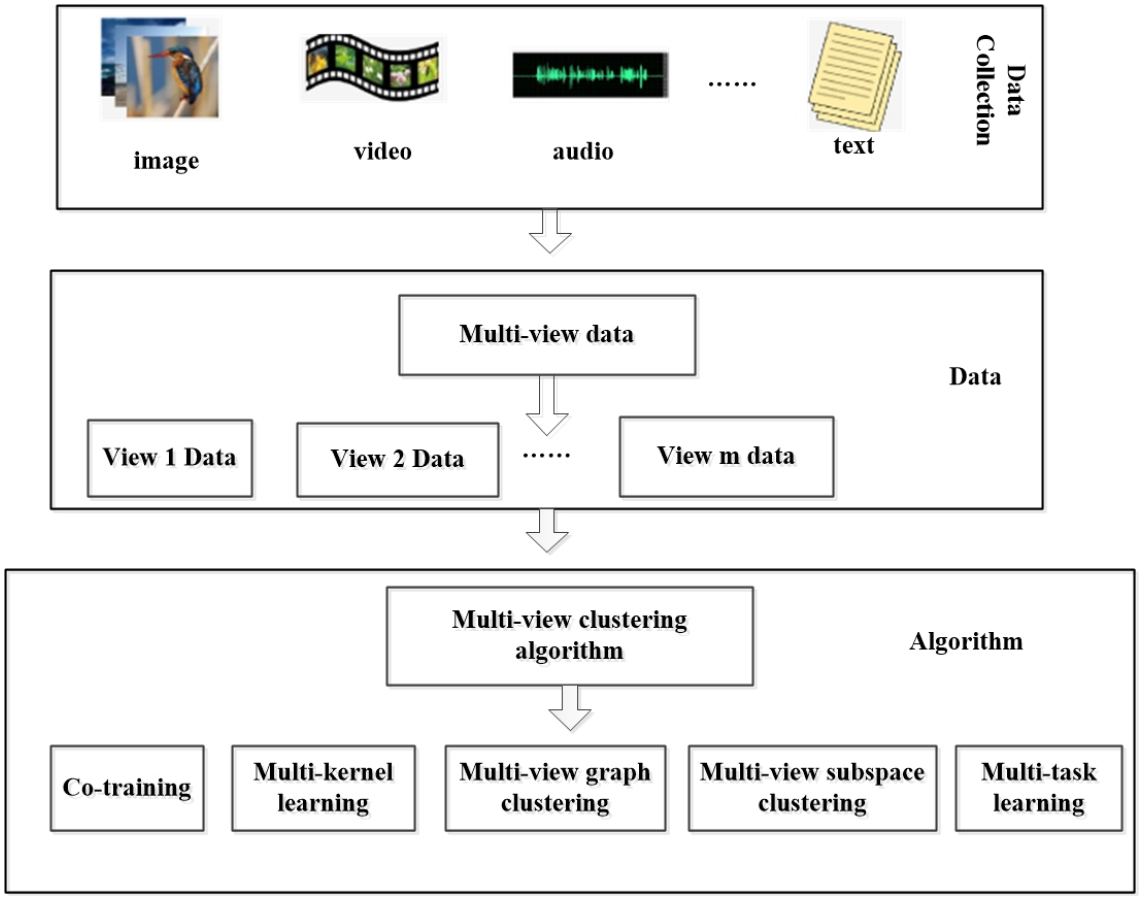


Figure 2.3: Components of multi-view data

Traditional machine learning algorithms can combine data from multiple views into one for learning when learning from multi-view data. However, such a simple concatenation does not effectively integrate information, which cannot improve learning performance from the rich multi-view information. Specifically, traditional clustering algorithms typically employ either feature selection from these data or handle each view independently in multi-view clustering. Nonetheless, these methods lack practical significance as they fail to fully leverage the complementary information from multiple data sources with diverse representations. Multi-view learning endeavors to integrate information from various views comprehensively, thus enhancing the performance of machine learning tasks.

Therefore, multi-view clustering algorithms constitute a prominent focus of contemporary research. These multi-view algorithms hold significant importance and categorically fall into five distinct types. The subsequent section will elaborate on

these five multi-view clustering algorithms.

2.3 Multi-view Clustering

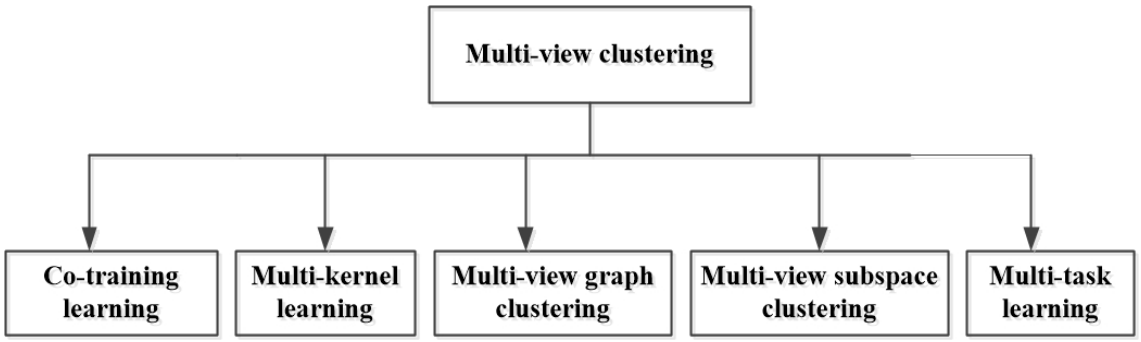


Figure 2.4: Five classifications of multi-view clustering

The following will analyse the advantages and disadvantages of these five algorithms: co-training, multi-kernel learning, multi-view graph clustering, multi-view subspace clustering, and multi-task learning:

Co-training effectively utilizes the compatibility and complementarity of multiple views. If the data consists of conditionally independent views, it implies that each view possesses sufficient information to produce the optimal learner. At the same time, conditional independence refers to the independence of two views under given class labelling conditions. However, collaborative training can only be used for two views and cannot handle multiple s.

Although multi-kernel learning has performed well in solving some multi-view datasets, efficiency is a bottleneck in developing multi-kernel. Firstly, multi-kernel learning algorithms are unsuitable for processing large-scale data due to their need to calculate the kernel combination coefficients for each kernel matrix, necessitating the involvement of multiple kernels in the operation and increased computational demands. If the data number is too large, the dimensionality of the kernel matrix is large, occupying a large amount of memory space. Secondly, it is very time-consuming in terms of time. The high time and spatial complexity are why multi-kernel learning cannot be

widely applied.

Multi-task learning is an essential technology in machine learning and data mining, which helps improve the learning ability of all tasks by utilizing the relevant knowledge of multiple tasks. Multi-task learning improves the learning performance of each task by learning common knowledge between these two tasks and transferring it between them. Research has shown that learning multiple tasks simultaneously yields better performance than learning each task individually. Multi-task learning has the following advantages:

1. Multiple tasks share a algorithm, reducing memory usage;

2. Calculating results from multiple tasks at once increases reasoning speed;

3. Associated tasks can enhance the advantages of various views by sharing information and complementing each other.

However, there are also the following drawbacks:

1. The convergence speed varies for different tasks, with simple tasks having a faster convergence speed and complex tasks having a slower convergence speed.

2. The inconsistent update direction of different tasks leads to problems with algorithm parameters and negative transfer between tasks.

The graph-based multi-view clustering algorithm does not require specifying the number of clusters parameter and other parameters that describe the number of clusters. It has an increased applicability and a clear clustering centre point, allowing data to be asymmetric. The data has extensive applicability but is not sensitive to initial values. Multiple executions of the clustering algorithm result in identical results, high algorithm complexity, and often take a long time to calculate. This will consume much time,

especially when running under massive data. Spectral graph clustering algorithms are easy to understand and can handle large-scale datasets with low time and spatial complexity. They can cluster on arbitrarily shaped sample spaces and converge to the global optimal solution. However, when the dataset is too large, the results are prone to local optima and are very sensitive to noise and outliers. Spectral clustering is also very sensitive to changes in similarity graphs and the selection of clustering parameters.

Multi-view subspace learning effectively addresses the challenge of high-dimensionality in multi-view data. However, it fails to harness the distinct characteristics among views to effectively learn a high-quality shared coefficient matrix or enhance the low rank of shared coefficients.

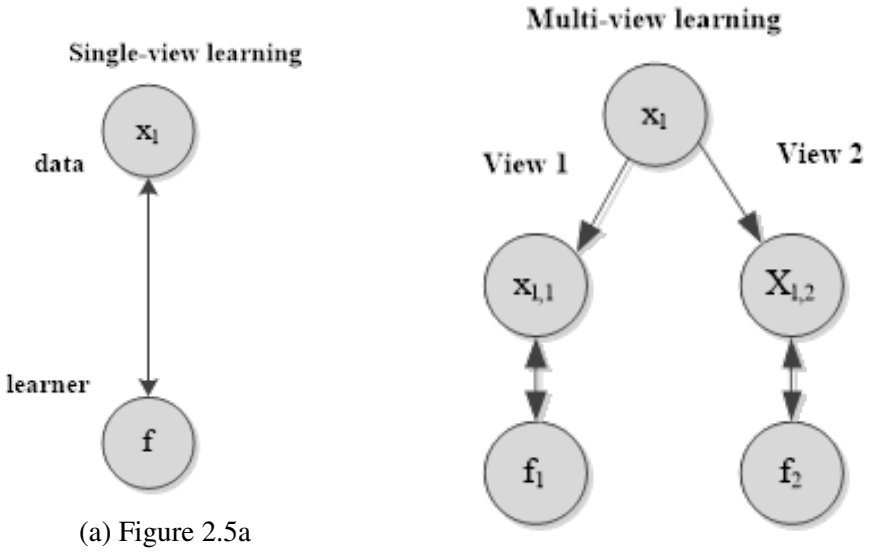
Multi-view clustering, a crucial data mining method, enhances clustering accuracy and robustness by integrating information from various views. The following subsection provides a detailed introduction to five commonly used multi-view clustering algorithms:

2.3.1 Co-training Algorithms

When dealing with -scale datasets containing multiple features, it becomes challenging to train robust classifiers independently on each view because a significant portion of the data remains unlabelled. This limitation hampers the development of classifiers with reliable generalization abilities. Therefore, collaborative training is adopted. Collaborative training algorithms were studied to address the consistency across multiple views (Predd et al., 2009). This approach maximizes mutual agreement among all views, leading to the broadest consensus.

Let's consider a dataset with two redundant views, namely view1 (X_1) and view2 (X_2). An example from this dataset can be represented as (x_1, x_2), where x_1 is the eigenvector of x in the X_1 view, and x_2 is the eigenvector of x in the X_2 view. Let f

denote the objective function in the example space X, as illustrated in Figure 2.5.



(a) Figure 2.5a

(b) Figure 2.5b

Figure 2.5: View data

Co-training represents a frequently employed technique within semi-supervised learning (X. Zhao et al., 2014). Figure 2.5(b) assumes that data samples can be represented by two conditionally independent features, x_1 and x_2. Two predictors, f_1 and f_2 assign a class label Y ($f : X \rightarrow Y$) to each instance X, leveraging a small pool of labelled data for training each feature. The two predictors assign labels to a r pool unlabelled data, incorporating the most likely subset of samples each predictor suggests into the labelled data pool. The predictors are then iteratively retrained and applied to the remaining unlabelled data. Collaborative training involves learning two distinct predictors, f_1 and f_2, which exhibit consistency on the unlabelled data across different views (X. Zhao et al., 2014). The co-training algorithm is introduced as follows (Blum & Mitchell, 1998):

Figure 2.6 (Y. Yang & Wang, 2018) illustrates the overall structure of the collaborative learning-based multi-view clustering algorithm. Two trainers, Trainer 1 and Trainer 2, engage in interactive and iterative training. They exchange their prior knowledge or learned information to maximize consistency between them.