Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

A Survey of Multiple Kernel Graph Clustering Algorithms

Zihao Li, Wenjing Chu

School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu 233030, China

Abstract: In recent years, algorithms based on multi-kernel learning and graph clustering have garnered significant attention in the field of data mining. This article reviews four advanced multi-kernel graph clustering algorithms: nearest neighbor linear kernel weighted multi-kernel graph clustering, multi-kernel graph clustering based on simultaneous global and local structure preservation, multi-kernel graph clustering based on direct consensus relationship graph learning, and custom non-negative matrix factorization multi-kernel graph tensor clustering. These algorithms have their own characteristics and, by combining multi-kernel learning, graph clustering, and other advanced technologies, they improve the accuracy and efficiency of clustering. However, they also face challenges such as high computational complexity and parameter settings, necessitating further research and optimization. These algorithms have wide application prospects in fields such as image recognition, text classification, and bioinformatics, providing new solutions for handling complex data structures.

Keywords: Multi-kernel learning, Graph clustering, Nearest neighbor linear kernel.

1. Introduction

With the rapid development of industries such as the Internet, video surveillance, military detection, medical imaging, and Earth remote sensing, acquiring complex data formats like text, images, videos, and audio has become effortless. These diverse data sources pose new challenges to scholars in the field of machine learning. How to effectively manage this data and extract valuable information from it has emerged as a crucial issue in machine learning, pattern recognition, data mining, and artificial intelligence. Clustering, as a vital data analysis method[1], plays an indispensable role.

The core objective of clustering is to partition the dataset into multiple meaningful groups or clusters composed of similar objects based on the representation information of the data and the similarity between them. This partitioning helps eliminate redundancy and useless information in the data, reduces uncertainty, and enhances the consistency within groups and the distinctiveness between groups, thereby providing a solid foundation for subsequent data processing or analysis tasks. It is worth noting that clustering is an unsupervised learning method, meaning that during the clustering process, we do not know the labels or affiliations of each data point. This characteristic of not requiring labeled training data gives clustering methods broad prospects for application.

In fact, clustering has played an important role in various fields such as information retrieval, biological data analysis, multimedia data analysis, feature representation, data cleaning, medical diagnosis, and anomaly detection, driving the rapid development of related theories and methods(as shown in Figure 1). However, with the increasing diversification of information collection methods and data types, the clustering problem of complex data has become more severe. This greatly limits the application of clustering-based algorithms in solving practical problems in production, life, and the national economy. Therefore, clustering for complex data has become an urgent frontier scientific problem. Clustering plays a crucial role in fields such as machine learning, artificial intelligence, and image processing.

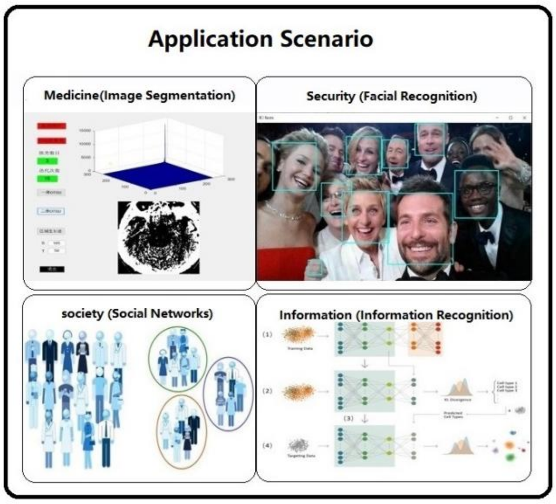

Figure 1. Examples of Clustering Applications

2. Multi-kernel Clustering

Over the past few decades, researchers have proposed many powerful clustering algorithms, which have achieved significant results on single-view data. However, when facing multi-view data, these algorithms often fall short. Even attempts to connect all views into one whole and then apply state-of-the-art clustering techniques on this single view often fail to achieve satisfactory clustering results. This is because each view carries its unique statistical characteristics and information, and simply merging them may not effectively utilize this information, leading to information loss and confusion.

The Multi-Kernel Clustering algorithm (MKC) has emerged to address this challenge, specifically designed for multi-view data to fully exploit the diversity and complementarity among different views. By considering the characteristics of multiple views, multi-kernel clustering can more comprehensively and deeply explore the structure and patterns in the data.

As an advanced clustering technique, multi-kernel clustering has attracted a great deal of attention and research from scholars in recent years. Due to the inherent connections and differences among different views in multi-view data, how to fully and reasonably utilize the information in these views becomes the key to improving the performance of multi-kernel learning.

To better utilize the information in multi-view data, multi-kernel clustering algorithms typically follow two core principles [1]:

1、Consistency Principle: This principle emphasizes that multiple kernels should maintain consistency and collaboration during the clustering process. In other words, although different kernels may have different characteristics and perspectives, they should reach some consensus on the clustering results, ensuring that samples belonging to the same category are correctly assigned to the same cluster across all kernels.

2、Complementarity Principle: This principle suggests that each kernel may contain information or knowledge that other kernels do not possess. Therefore, by considering the information from multiple kernels comprehensively, multi-kernel clustering algorithms can obtain more comprehensive and accurate clustering results.

This article mainly introduces four novel multi-kernel clustering algorithms, namely the nearest neighbor linear kernel weighted multi-kernel graph clustering algorithm, the multi-kernel graph clustering algorithm based on simultaneously preserving global and local structures, the multi-kernel graph clustering algorithm based on direct consensus relationship graph learning, and the multi-kernel graph tensor clustering algorithm customized for non-negative matrix factorization.

2.1. Nearest Neighbor Linear Kernel Weighted Multi-Kernel Graph Clustering

The nearest neighbor linear kernel weighted multi-kernel graph clustering is an advanced clustering algorithm that combines the advantages of multi-kernel learning and graph clustering to enhance clustering performance through the weighted integration of linear kernels based on nearest neighbor relationships.

This algorithm addresses the limitations of traditional clustering algorithms in handling complex data structures by introducing the concept of multi-kernel learning. Multi-kernel learning allows the algorithm to utilize multiple different kernel functions to capture various features of the data, thereby improving clustering accuracy.

Moreover, the core idea of the algorithm lies in the nearest neighbor linear kernel weighting. It calculates the similarity between different kernel functions and assigns them different weights to construct a weighted multi-kernel matrix. This approach maximizes the advantages of different kernel functions while reducing the impact of noise and redundant information[2].

Furthermore, the algorithm incorporates the idea of graph clustering, which is a clustering method based on graph theory that discovers cluster structures in data by constructing a graph model. In the nearest neighbor linear kernel weighted multi-kernel graph clustering, the algorithm first constructs a consensus relationship graph and then partitions it using graph clustering algorithms to obtain the final clustering result. It leverages both global and local structural information of the data while considering the complementarity between different kernel functions, demonstrating excellent performance in handling datasets with complex structures and features. However, the algorithm's computational complexity is relatively high, requiring significant computational resources.

In terms of applications, the nearest neighbor linear kernel weighted multi-kernel graph clustering algorithm has been widely used in fields such as image recognition, text classification, and bioinformatics. In these domains, the algorithm effectively handles large-scale high-dimensional data, uncovering the intrinsic structures and patterns within the data. Future research directions include optimizing the algorithm's computational efficiency, reducing computational complexity, designing more effective kernel function selection strategies to adapt to different types of datasets, and integrating other advanced machine learning techniques to further enhance clustering performance and accuracy.

2.2. The multi-kernel graph clustering algorithm based on simultaneously preserving global and local structures

The Structure Preserving Multiple Kernel Clustering algorithm is a graph clustering approach aimed at improving the quality of relational graphs by considering both global and local structure preservation in the kernel space of the data.

Specifically, the SPMKC algorithm proposes a strategy based on heat kernel and nearest neighbor kernel weighting to learn the optimal consensus kernel. This strategy helps the algorithm to accurately identify the intrinsic structure and features of the data during the clustering process. Additionally, the algorithm emphasizes preserving the global and local structures of the data, meaning that it attempts to maintain the relative positions and relationships between data points as much as possible during clustering to better reflect the true distribution and characteristics of the data. Through this approach, the SPMKC algorithm achieves better performance in multi-kernel graph clustering, enhancing clustering accuracy and stability.[3] It holds significant potential for practical applications, particularly in dealing with complex and large-scale datasets.

It's worth noting that while the SPMKC algorithm excels in preserving global and local structures, its performance is still influenced by various factors such as data distribution, noise interference, and algorithm parameter settings. Therefore, in practical applications, adjustments and optimizations to the algorithm may be necessary to achieve the best clustering results based on specific circumstances.

2.3. The multiple kernel graph clustering algorithm based on direct consensus relationship graph learning.

The Consensus Relationship Graph Learning-based Multiple Kernel Graph Clustering Algorithm is an innovative clustering approach designed to address efficiency issues of traditional clustering algorithms when dealing with large-scale high-dimensional data. This algorithm is based on the research idea of pure graph multiple kernel graph clustering, aiming to effectively discover the true cluster structure of data by directly learning consensus relationship graphs.

CAGL simplifies the algorithm process and reduces computational complexity by optimizing objectives, enabling it to maintain high efficiency when processing large-scale datasets. Through an iterative self-adjusting Laplacian rank constraint weight method, CAGL ensures precise connectivity components in the consensus relationship graph. This approach aids in better identifying the genuine cluster structure of data, thus improving clustering accuracy. The algorithm also adopts a nearest k-sparse strategy to reduce redundant edges in the relationship graph. This not only lowers computational complexity but also enhances inter-class sparsity, resulting in clearer clustering outcomes.

A self-weighted graph fusion strategy is proposed to integrate complementary and consensus information from different candidate relationship graphs, avoiding the high computational complexity associated with quadratic programming. This feature enables the algorithm to conserve computational resources while ensuring accuracy.

In application, the Consensus Relationship Graph Learning-based Multiple Kernel Graph Clustering Algorithm can be employed in various clustering scenarios, particularly for large-scale high-dimensional datasets with complex structures and features, showcasing superior performance. Therefore, this algorithm represents an efficient and accurate clustering method, leveraging optimization objectives, iterative self-adjustment of weights, nearest k-sparse strategy, and self-weighted graph fusion strategy to enhance clustering accuracy and efficiency.

2.4. The Multi-Kernel Graph Tensor Clustering Algorithm Customized by Non-negative Matrix Factorization

The custom Non-negative Matrix Factorization based multi-kernel graph tensor clustering algorithm is an advanced clustering technique that combines the principles of NMF and multi-kernel graph tensor clustering. It aims to enhance clustering effectiveness and performance by leveraging the characteristics of NMF and the advantages of multi-kernel graph tensor clustering[4].

NMF is a matrix factorization method that requires all components of the decomposition to be non-negative. This method enables nonlinear dimensionality reduction while preserving the non-negativity of the data. By decomposing the original data into basis and coefficient matrices, NMF reveals the intrinsic structure and features of the data. Multi-kernel graph tensor clustering, on the other hand, is a clustering algorithm based on multi-kernel learning and graph tensor theory. It captures different data characteristics using multiple kernel functions and constructs a graph model of the data based on tensor theory to discover cluster structures. This approach effectively utilizes both global and local structural information of the data, thereby improving clustering accuracy[5].

The custom NMF-based multi-kernel graph tensor clustering algorithm integrates NMF principles into multi-kernel graph tensor clustering through customization. Specifically, the algorithm preprocesses the original data using NMF to obtain basis and coefficient matrices. Then, it constructs a multi-kernel graph tensor model based on these matrices and performs clustering using multi-kernel graph tensor clustering methods. This approach fully utilizes the advantages of NMF in data representation and feature extraction, while combining the capability of multi-kernel graph tensor clustering to discover data cluster structures, thus enhancing overall clustering performance.

The algorithm's strengths lie in its ability to fully utilize the non-negativity, global and local structural information of the data, as well as the complementary nature of multiple kernel functions. This makes it particularly effective in handling datasets with complex structures and features. However, the algorithm's computational complexity is relatively high, requiring significant computational resources.

In terms of applications, the custom NMF-based multi-kernel graph tensor clustering algorithm can be applied to various clustering scenarios, especially for large-scale high-dimensional datasets with complex structures and features, where it demonstrates superior performance.

Future research directions include further optimizing the algorithm's computational efficiency and reducing its complexity, designing more effective kernel function selection strategies to adapt to different types of datasets, and integrating other advanced machine learning techniques to further enhance clustering performance and accuracy.

In summary, the custom NMF-based multi-kernel graph tensor clustering algorithm is an advanced clustering technique that combines NMF and multi-kernel graph tensor clustering methods. By customizing NMF principles into multi-kernel graph tensor clustering, it improves clustering effectiveness and performance. However, the algorithm still faces some challenges and requires further research.

3. Conclusion

The core idea of the nearest neighbor linear kernel weighted multi-kernel graph clustering algorithm is to introduce the concept of neighborhood based on the traditional linear kernel weighted research approach. It treats the kernel generated by linear weighting as a pseudo-consensus kernel and searches for the optimal consensus kernel within the neighborhood range of the pseudo kernel. By considering neighborhood information, the algorithm can more accurately capture the local structure of the data, thereby improving the accuracy of clustering. However, this algorithm may overlook the consideration of the global structure of the data, leading to situations where the advantages of multi-kernel information cannot be fully utilized. This drawback can be addressed by a multi-kernel graph clustering algorithm based on simultaneously maintaining global and local structures. This algorithm considers both the global and local structural information of the data during the clustering process, aiming to preserve the structural information of the original data in its original space. By simultaneously considering both global and local structures, the algorithm can more comprehensively reflect the characteristics of the data, thereby improving the stability and accuracy of clustering. It may face the problem of high computational complexity when dealing with large-scale datasets, requiring a trade-off between computational efficiency and clustering effectiveness. The non-negative matrix decomposition customized multi-kernel graph tensor clustering algorithm utilizes the non-negative matrix decomposition method to decompose the multi-kernel graph tensor, thereby obtaining clustering results. Non-negative matrix decomposition has the dual functions of dimensionality reduction and clustering, which can simultaneously achieve data dimensionality reduction and clustering, improving the efficiency of the algorithm. However, it may be necessary to adjust the parameters and strategies of non-negative matrix decomposition according to different datasets and clustering requirements to achieve the best clustering results.

Overall, these algorithms have their own strengths in the field of multi-kernel graph clustering, suitable for different scenarios and requirements. In practical applications, it is necessary to choose the appropriate algorithm based on the specific dataset, computational resources, and clustering objectives. Furthermore, as technology continues to evolve, these algorithms will continue to be optimized and improved to adapt to a wider range of application scenarios.

Acknowledgment

This work was supported by the Undergraduate Research Innovation Project of Anhui University of Finance and Economics (Grant NO. XSKY23156). The authors declare that they have no conflicts of interest.

References

[1] Zhenwen Ren. Research of Multiple Kernel GraphClustering. Nanjing university of science and technology, 2021.

[2] ZHANG Xiaoqian,WANG Xiao. Image Segmentation Algorithm Based on Weighted Multi-Kernel Subspace Clustering. Journal of Beijing University of Posts and Telecommunications, 2023, 46(03): 78-83.

[3] ZHAO Lifeng, LI Xin ,WANG Dong. Clustering Algorithm Based on Multiple Kernel SVM. Periodical of Ocean University of China,2009, 39(05):1047-1050

[4] Xia Dongxue ,Yang Yan ,Wang Hao. Late Fusion Multi-View Clustering Based on Local Multi-Kernel Learning. Journal of Computer Research and Development, 2020, 57(08): 1627-1638.

[5] HE Xue-mei. A Survey of Multi-view Clustering Algorithms. Software Guide, 2019, 18(4): 79-81, 86.