WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n2.09
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Theory and Algorithms of Multi-Kernel Graph Clusterings

Zihao Li, Wenjing Chu

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

ABSTRACT

This article introduces the increasingly prominent importance of data analysis and mining due to the explosive growth of data in the information age, particularly in the analysis and mining of graph data. The characteristics of graph data lie in the complex connections between nodes, making its analysis and mining a hot research topic. Traditional clustering algorithms have limitations when dealing with non-linearly separable data, leading to the emergence of multi-kernel graph clustering algorithms. These algorithms utilize multiple kernel functions to compute the similarity between samples in a high-dimensional feature space, thereby better capturing data features and providing more accurate clustering results. The article primarily investigates the principles, algorithms, and applications of various multi-kernel graph clustering algorithms, emphasizing their advantages in handling non-linearly separable data and offering more accurate clustering results. It suggests that further research into these algorithms will enhance clustering algorithm performance and achieve better results in practical applications.

KEYWORDS

Multi-kernel learning, Graph clustering, Data mining

1. INTRODUCTION

With the advent of the information age, the explosive growth of data has made data analysis and mining increasingly important. Graph data, as an important form of data, has been widely used in fields such as social networks, bioinformatics, and image processing. The characteristic of graph data is the complex connectivity between nodes, which often contains rich information and hidden patterns. Therefore, the analysis and mining of graph data have become a hot research direction. In the field of machine learning and data mining, clustering is a commonly used unsupervised learning method for grouping similar data points into clusters. Traditional clustering algorithms such as k-means and hierarchical clustering perform well when dealing with linearly separable data, but they have limitations when handling nonlinearly separable data. To address this issue, multi-kernel graph clustering algorithms have emerged.

Multi-kernel graph clustering algorithms are a class of algorithms that use multiple kernel functions for clustering. They map the data to a high-dimensional feature space and use different kernel functions to calculate the similarity between samples in this space, thereby capturing richer data features. The advantage of multi-kernel graph clustering algorithms is that they can better handle nonlinearly separable datasets and provide more accurate clustering results.

In recent years, multi-kernel graph clustering algorithms have attracted widespread research interest in the fields of machine learning and data mining. Researchers have proposed many different multi-kernel graph clustering algorithms to adapt to different data types and problem requirements. These

algorithms mainly include spectral clustering-based multi-kernel graph clustering, kernel k-means-based multi-kernel graph clustering, and kernel spectral clustering-based multi-kernel graph clustering, among others.

Spectral clustering-based multi-kernel graph clustering algorithms are the most common and classical type of method. It constructs the similarity matrix of the data and performs clustering using the eigenvectors of this matrix. Different kernel functions can be used to calculate the similarity matrix, capturing different data features. Commonly used kernel functions include Gaussian kernel, polynomial kernel, and Laplacian kernel, among others.

Kernel k-means-based multi-kernel graph clustering algorithms perform clustering within the framework of multi-kernel learning. It achieves clustering by alternately optimizing kernel weights and sample cluster assignments. Kernel weights represent the importance of different kernel functions, while sample cluster assignments represent which cluster each sample point belongs to. By iteratively optimizing these two variables, the optimal clustering result can be obtained.

Kernel spectral clustering-based multi-kernel graph clustering algorithms combine the ideas of spectral clustering and multi-kernel learning. It constructs a weighted graph and performs clustering using the eigenvectors of the graph's Laplacian matrix. Different kernel functions can be used to calculate the edge weights of the weighted graph, thereby affecting the clustering result.

Multi-kernel graph clustering algorithms are a powerful clustering method with better adaptability and flexibility. They can effectively handle nonlinearly separable data and provide more accurate clustering results. In the following articles, we will introduce the principles, algorithms, and applications of multi-kernel graph clustering algorithms in detail, and conduct in-depth research and analysis on them. Through the study of multi-kernel graph clustering algorithms, we can further improve the performance of clustering algorithms and achieve better results in practical applications.

2. MULTI-KERNEL CLUSTERING

2.1. Spectral clustering based multi-kernel graph clustering algorithm

Spectral clustering is widely favored by scholars due to its excellent performance and comprehensive mathematical theory. It constructs a similarity matrix of the data and utilizes the eigenvectors of this matrix for clustering. Different kernel functions can be used to calculate the similarity matrix, capturing various data features. Common kernel functions include Gaussian kernel, polynomial kernel, and Laplacian kernel, among others. The specific steps include: Firstly, constructing a relationship graph of the original data using linear subspace learning methods, where each point in the graph reflects the similarity relationship between pairs of data; Secondly, performing spectral clustering on the learned relationship graph to assign data points to corresponding disjoint subspaces and obtain clustering results.(as shown in Figure 1) It can be seen that the quality of the relationship graph directly determines the clustering results. Therefore, how to construct a high-quality relationship graph is currently a key research focus. In the real world, data often exhibit nonlinear distributions, leading to poor clustering performance of linear subspace learning methods in practical applications. Additionally, real-world data often contain outliers or noise, which can damage the true structural distribution of data in subspaces, resulting in incorrect clustering results. Based on this, existing spectral clustering algorithms urgently need to address these two challenging issues: first, how to avoid the influence of noise or outliers in the real world and learn high-quality relationship graphs from noisy data; second, how to enable linear subspace algorithms to learn high-quality relationship graphs even on nonlinear data. To address the first issue, clustering algorithms based on Low-Rank Representation and Sparse Representation have been widely researched. Among them, LRR constrains the relationship graph with nuclear norm to induce low-rankness, while SR constrains the relationship graph with $\ell 1$ norm to induce sparsity. Therefore, theoretically combining the two can simultaneously preserve the low-rank and sparse structural information of the relationship graph,

effectively mining the low-rank and sparse properties of the relationship graph in subspaces, which is very effective for learning high-quality relationship graphs[1].

Figure 1. The steps of the Spectral clustering algorithm

2.2. The k-means based multi-kernel graph clustering algorithm

The multi-kernel K-means based graph clustering method conducts clustering within the framework of multiple kernel learning. It achieves clustering by iteratively optimizing kernel weights and sample cluster assignments. Kernel weights represent the importance of different kernel functions, while sample cluster assignments indicate which cluster each sample point belongs to. By iteratively optimizing these two variables, the optimal clustering results can be obtained[2].

The multi-kernel K-means based graph clustering method is an algorithm that combines multiple kernel learning and K-means clustering. It aims to achieve clustering by optimizing kernel weights and sample cluster assignments. Below are the main steps of the multi-kernel K-means based graph clustering method:

Firstly, given a dataset, different kernel functions are selected to compute the similarity between each pair of samples. These similarities form a kernel matrix, where each element represents the similarity between two samples. Initialize a weight for each kernel function and randomly assign a cluster to each sample. In this step, the clustering results are continuously improved by iteratively optimizing kernel weights and sample cluster assignments[3]. Fix the sample cluster assignments and update the weights of each kernel function. A common approach is to minimize the objective function of kernel K-means, which considers the distance between each sample and the centroid of its cluster. Fix the kernel weights and update the clusters to which each sample belongs. Conventional K-means clustering algorithms can be used to accomplish this step. Stop iterating and output the final clustering results when the kernel weights and sample cluster assignments no longer change significantly.

The advantage of the multi-kernel K-means based graph clustering method lies in its ability to better capture the different features of various kernel functions, thereby improving the accuracy of clustering. Additionally, it exhibits good robustness and can handle noise and outliers(as shown in Figure 2).

However, the multi-kernel K-means based graph clustering method also faces some challenges. One of them is determining the strategy for selecting appropriate kernel functions and kernel weights. Additionally, the method may encounter high computational complexity when dealing with large-scale data.

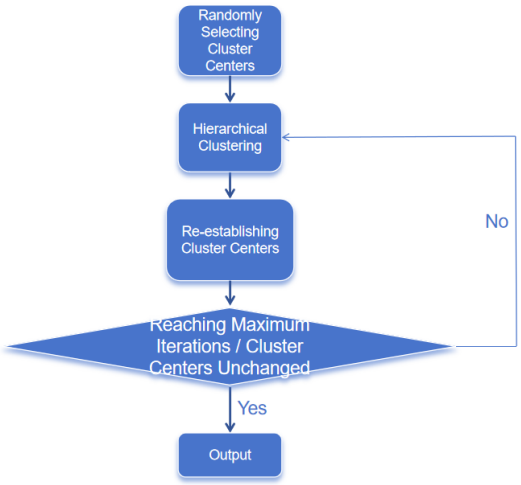

Figure 2. The steps of the K-means algorithm

Overall, the multi-kernel K-means based graph clustering method is a powerful clustering algorithm that can provide better clustering results when dealing with nonlinearly separable data. By carefully selecting kernel functions and optimizing kernel weights, the performance of the clustering algorithm can be further improved[4].

2.3. The Kernel spectral clustering based multi-kernel graph clustering algorithm.

This method combines the ideas of spectral clustering and multi-kernel learning. It achieves clustering tasks by constructing a weighted graph and utilizing the eigenvectors of the graph's Laplacian matrix for clustering. Different kernel functions can be used to compute the edge weights of the weighted graph, thereby affecting the clustering results. Multi-kernel graph clustering algorithms are a powerful class of clustering methods with better adaptability and flexibility. They can effectively handle nonlinearly separable data and provide more accurate clustering results. In the following article, we will provide a detailed introduction to the principles, methods, and applications of multi-kernel graph clustering algorithms, and conduct in-depth research and analysis on them. Through studying multi-kernel graph clustering algorithms, we can further improve the performance of clustering algorithms and achieve better results in practical applications.

The multi-kernel graph clustering method based on kernel spectral clustering combines the ideas of kernel spectral clustering and multi-kernel learning, aiming to achieve clustering tasks by optimizing kernel weights and the spectral clustering process. Below is a detailed description of this method:

Firstly, based on the given dataset, different kernel functions are selected to compute the similarity between each pair of samples. These similarities form a kernel matrix, where each element represents the similarity between two samples. Initialize a weight for each kernel function, which can typically be done using uniform weights or random initialization. Multiply the kernel matrix of each kernel function by the corresponding kernel weight to obtain a weighted kernel similarity matrix. Use the constructed weighted kernel graph for spectral clustering. This step usually includes the following steps: Calculate the Laplacian matrix: Calculate the unnormalized or normalized Laplacian matrix based on the adjacency matrix. Compute eigenvectors: Solve for the eigenvectors of the Laplacian matrix, typically selecting the eigenvectors corresponding to smaller eigenvalues as the new feature space. Perform clustering using the eigenvectors, typically clustering the eigenvectors using methods such as K-means to obtain the final sample clusters. On the basis of spectral clustering, optimize the kernel weights by minimizing the spectral clustering objective function, considering both the similarity between samples and the dissimilarity between clusters. Common optimization methods include gradient descent, EM algorithm, etc. Alternately perform spectral clustering and kernel weight optimization steps until convergence. You can usually set a maximum number of iterations or stop the iterations when the change in the optimization objective function is below a certain threshold.(as shown in Figure 3) When the kernel weights and spectral clustering results no longer change significantly, stop the iterations and output the final clustering results[5].

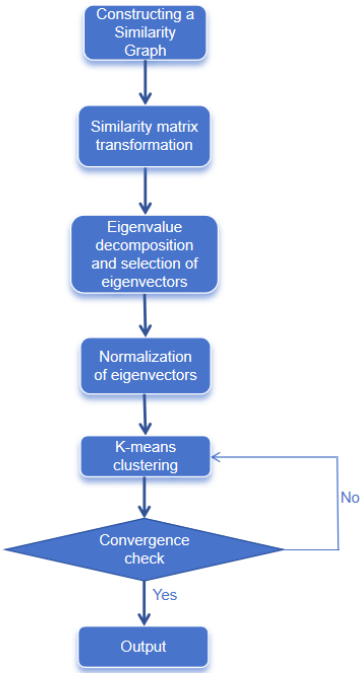

Figure 3. The steps of the Kernel spectral clustering algorithm

Through the above steps, the multi-kernel graph clustering method based on kernel spectral clustering can achieve good clustering results. This method fully utilizes the information from multiple kernel functions and combines the advantages of spectral clustering. It can effectively handle nonlinearly separable datasets and has good robustness.

3. CONCLUSION

Spectral clustering is effective in handling nonlinearly separable data and is applicable to clusters of various shapes. It maintains relatively low computational complexity when dealing with large-scale data. However, it is sensitive to parameter choices, such as constructing the similarity matrix and selecting the number of clusters. Performance may degrade in the presence of noise or outliers, and scalability to large datasets is limited.

Based on kernel K-means, the multi-kernel graph clustering method exhibits flexibility in handling nonlinearly separable data and automatically learns the weights of kernel functions, thereby enhancing clustering accuracy. By iteratively optimizing kernel weights and sample cluster assignments, it yields favorable clustering results. Nonetheless, it entails higher computational complexity for large-scale datasets, leading to longer execution times. Some parameter tuning may be necessary to achieve optimal results on certain datasets.

Multi-kernel graph clustering based on kernel spectral clustering combines the principles of spectral clustering and multi-kernel learning, effectively capturing the nonlinear characteristics of data and improving clustering accuracy. By leveraging different kernel functions to compute weighted graph edges, it adapts to diverse dataset features. However, this algorithm incurs higher computational complexity for large-scale datasets, necessitating additional computational resources. Manual adjustment of kernel function parameters may be required in some cases to obtain optimal results.

In summary, spectral clustering is suitable for clusters of various shapes and nonlinearly separable data but is sensitive to parameters. Kernel K-means-based multi-kernel graph clustering automatically learns kernel function weights but has higher computational complexity. Multi-kernel graph clustering based on kernel spectral clustering combines the advantages of spectral clustering and multi-kernel learning but entails higher computational complexity for large-scale datasets. The choice

of method tailored to specific problems and datasets requires considering factors such as algorithm performance, computational complexity, and data characteristics.

ACKNOWLEDGMENTS

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.

REFERENCE

[1] LIU Xiao-li, MOU Yi-hong. Semi-supervised Spectral Clustering Algorithm Based on Active Learning [J]. Journal of Gansu Normal Colleges, 2021, 26 (02): 41-45.

[2] Wang Sen, Liu Chen, Xing Shuaijie.Review on K-means Clustering Algorithm [J]. Journal of East China Jiaotong University, 2022, 39(05): 119-126

[3] Li Hengbo, Liu Jingchao, Wu Ketong. Image segmentation based on K-means algorithm [J]. Modern Computer, 2024, 30(02): 49-51+91.

[4] HAN Yu, WANG Qing, LIU Li'na .Optimization Research of K-means Clustering Algorithm [J]. SOFTWARE, 2023, 44(10): 58-61.

[5] WANG Weidong,LIU Bing,GUAN Hongjie. Spectral embedded clustering algorithm based on kernel function [J]. Journal of Computer Applications, 2015, 35(3): 761-765, 810.