Multi-dimensional Data Clustering Based on Network Community Detection Methods (Post-print)

Authors: Xingbin Wu, Guo Qiang, Zhang Linbing, Liang Yaozhou, Liu Jianguo

Date: 2018-12-13T00:00:00+00:00

Abstract

To address the issue of poor clustering performance of traditional clustering methods on multi-dimensional datasets, a method of network community detection is proposed and applied to multi-dimensional data clustering analysis. For a multi-dimensional dataset, feature extraction is first performed on the analysis objects to construct feature vectors for each object. The similarity between different feature vectors is measured by calculating the Pearson correlation coefficient, thereby constructing a similarity network. The Blondel algorithm is then employed to perform community detection on this network to achieve clustering. Experimental results demonstrate that this method can achieve favorable clustering results in multi-dimensional data clustering, with an accuracy rate of 92.5%, which outperforms the 75% accuracy of the k-means algorithm.

Full Text

Preamble

Vol. 37 No. 2
Application Research of Computers
ChinaXiv Partner Journal

Multidimensional Data Clustering Based on Network Community Detection Method

Wu Xingbin[1], Guo Qiang[1], Zhang Linbing[1], Liang Yaozhou[1], Liu Jianguo[2]†
(1. Research Center for Complex Systems Science, University of Shanghai for Science & Technology, Shanghai 200093, China;
2. Institute of Fintech, Shanghai University of Finance & Economics, Shanghai 200433, China)

Abstract: Since the traditional clustering method has an impoverished clustering effect on multidimensional data, this paper used the clustering method based on community detection of complex networks to achieve better results. Firstly, the method extracted the features of original data and formed the feature vectors of each object for a multidimensional data set. Then it measured the similarity between different feature vectors by calculating Pearson correlation coefficients and constructed a similarity network. Finally, it used the Blondel algorithm which detects the community of the network to achieve the clustering effect. Experimental results show that this method can get better clustering results in multidimensional data clustering with an accuracy rate of 92.5%, which is better than k-means algorithm with the accuracy rate of 75%.

Key words: clustering; multidimensional data; similarity; community detection

0 Introduction

How to analyze multidimensional data and mine the information hidden behind it has increasingly attracted widespread attention. Effective and accurate clustering is of great significance for identifying similar objects and mining the similarities and differences among different categories. Affected by the curse of dimensionality [1], traditional clustering methods perform well in low-dimensional data spaces but often fail to obtain good clustering results in high-dimensional data spaces. Finding effective clustering methods for multidimensional data has become an important direction in clustering research.

As a common technique in statistical analysis and a major task in data mining, clustering [2] has been extensively studied by scholars both domestically and internationally. Celebi et al. [3] investigated existing initialization methods to address the problem of K-means algorithm' s sensitivity to the initial positions of cluster centers. Kriegel et al. [4] studied density-based clustering, defining clusters as regions with higher density than the rest of the dataset, while objects in sparse regions are usually referred to as noise and boundary points. Tran et al. [5] proposed an improved DBSCAN algorithm to solve the problem of instability when conventional algorithms detect boundary objects of adjacent clusters. Ding et al. [6] proposed a novel iterative clustering framework to address the issue that differences between categories can mask differences between subclasses, which can further identify subtle differences after recognizing major categories and provide a comprehensive clustering trajectory. Zhang [7] proposed a new Pythagorean fuzzy clustering method that can solve clustering problems where criterion weights are not precisely given in advance.

In addition to seeking new clustering algorithms, the question of whether methods from different fields can be borrowed and applied to multidimensional data analysis to achieve better clustering results has prompted deeper thinking. After decades of development, complex network research has been widely applied

in numerous fields such as social sciences [8, 9], computer science [10], and biological sciences [11]. In 2002, Girvan and Newman pointed out that clustering characteristics are ubiquitous in complex networks [12], and in 2004 they proposed the GN algorithm for discovering these communities [13]. Since then, a large number of scholars have conducted extensive research on community detection problems in complex networks. Papadopoulos et al. first constructed the concept of community in the context of social media [14]. Xie et al. [15] summarized benchmarks for quality metrics of state-of-the-art overlapping community detection algorithms and found that different algorithms have different stability in networks with different overlapping densities. Community detection still has some unresolved issues, such as the lack of a unified definition of community itself and no universal standard for algorithm validation and performance comparison. Fortunato et al. [16] proposed a set of guidelines to address these issues, pointing out the advantages and disadvantages of existing popular algorithms and providing directions for the use of different algorithms. Yang et al. [17] proposed a community structure detection algorithm for directed networks based on multiple eigenvectors of the Laplacian matrix and simultaneously studied the evolution characteristics of new community members in dynamic networks [18].

The concept of community in networks is consistent with the concept of class in conventional clustering algorithms, so applying community detection algorithms to clustering analysis is feasible. This paper focuses on the problem that traditional clustering methods have poor clustering effects on multidimensional data and combines the idea of complex network community detection to apply network partitioning methods to multidimensional data clustering. The results show that compared with traditional clustering methods, network partitioning can achieve better clustering performance.

1 Related Theory

For the analysis of multidimensional datasets, this paper designed a processing pipeline. First, the original dataset is preprocessed to remove abnormal data such as outliers and missing values. Then, according to analysis requirements and actual scenarios, features of the research objects are extracted. Subsequently, the correlation coefficients between features of different objects are calculated to construct a similarity network. The Blondel algorithm is then applied to this network for community detection to achieve clustering effects. Finally, the clustering results can be further analyzed. The experimental process is shown in Figure 1 [Figure 1: see original paper].

1.1 Network Construction

Network nodes represent clustering objects, and edges represent correlations between objects. The Pearson correlation coefficient between any two objects i and j is defined by Equation (1). If the Pearson correlation coefficient exceeds the threshold , we consider that there is an edge between nodes i and j, where

is the threshold point. Selecting an appropriate threshold can construct a network with obvious topological structure, which is beneficial for subsequent research.

1.2 Blondel Community Detection Algorithm

Newman et al. first proposed the concept of modularity in 2004, which is a metric for measuring the quality of community detection, as shown in Equation (2). For the network under study, A_ij represents the adjacency matrix of the network, k_i represents the degree of node i, m is the number of edges in the network, and c_i represents the community to which node i belongs. (c_i, c_j) equals 1 if nodes i and j are in the same community, otherwise it is 0. The Q value ranges between 0 and 1, and a larger Q value indicates more effective community structure.

Blondel et al. [19] proposed a fast clustering algorithm based on modularity in 2008. Its main goal is to continuously partition communities so that the modularity of the entire network after partitioning keeps increasing. The greater the modularity of the partitioned network, the better the community detection effect. The literature proposes that when node i is partitioned into community C, the modularity gain of community C is calculated by Equation (3), where Σ_in is the sum of internal connection weights within community C, Σ_tot is the sum of all connection weights pointing to nodes in community C, and k_i,in is the sum of connection weights between node i and other nodes in the community.

The algorithm is described as follows:
a) Treat each node in the network as a separate community;
b) For each node, attempt to partition it into the community where its adjacent nodes are located, calculate the modularity at this time, and determine whether the difference in modularity ΔQ before and after partitioning is positive. If positive, accept this partitioning; if not, abandon it;
c) Repeat step b) until the modularity no longer increases after community partitioning;
d) Treat each community obtained in step c) as a new node, construct a new network, and repeat steps b) and c) until the community structure no longer changes.

2.1 Experimental Data

This paper adopted data from 133 drivers and 55 vehicles on a domestic bus route in September 2017, conducting clustering analysis research on the bus drivers of this route. The dataset contains 17 fields including record date, record time, vehicle ID, driver ID, location longitude, location latitude, CAN speed, etc.

2.2 Data Cleaning

First, the original data was cleaned. For missing data that occurred during the actual data collection process, this paper adopted the hot deck imputation method [20], i.e., the nearest completion method, based on actual business analysis. This paper conducted logical error detection on the original data. For example, in some trajectory segments, the speed data of some vehicles exceeded the normal range and reached 120 km/h. The processing method for this unconventional data was to delete the trajectory segment. For some fields with substantial data missing, such as the driver brake pedal opening field with a missing rate close to 100%, this paper conducted abandonment processing. For some fields with outliers, such as motor speed reaching an abnormal value of 16,383 rpm, the outliers were directly replaced with blank values, which were then imputed by nearest completion.

2.3 Feature Acquisition

To make driver behavior features comparable, this paper defined a specific analysis scenario: a driver driving a bus from the starting point to the endpoint constitutes a complete trip record, which serves as the benchmark for calculating each driver' s behavior features. First, the cleaned data was segmented according to the trip scenario, and driver behavior features for each trip were extracted from the sliced data. Based on valid data and scenario inference, this paper selected nine features: mean absolute acceleration; standard deviation of absolute acceleration; mean speed; standard deviation of speed; electronic brake usage probability; foot brake usage probability; mean throttle pedal percentage; standard deviation of throttle pedal percentage; and median speed. Regarding the positive and negative values of acceleration caused by vehicle acceleration and deceleration, this paper took the absolute value of acceleration.

2.4 Similarity Network Modeling

A driver' s trip on a bus route can be represented as an n-dimensional vector D_i = {x_i1, x_i2, ···, x_in}, where i represents the ith trip of the driver, and x_i1, x_i2, ···, x_in represent n behavioral feature attributes related to the driver' s driving during the ith trip. The network is constructed based on the similarity of behavioral features of different trips. This paper adopts the Pearson correlation coefficient to measure the feature similarity between different drivers, as shown in Equation (1), where D_i and D_j are the driving behavior feature attribute vectors of the ith and jth trips, respectively. If the driving behavior features of two trips have certain similarity in each component feature, an edge is constructed between these two driving behavior feature vectors.

Based on Equation (1) to calculate similarity, a threshold of 0.15 was set. When the similarity between driving behavior feature vectors of different trips exceeds this threshold, an edge is established. The final network contains 879 nodes and 386,760 edges, with an average node degree of 440.

2.5 Clustering Results and Accuracy Measurement

Based on the network constructed using Pearson correlation coefficients, the Blondel algorithm was applied for community detection, yielding three communities. Class 1 contains 679 driving behaviors from 101 different drivers, Class 2 contains 199 driving behaviors from 40 different drivers, and Class 3 contains 1 driving behavior from 1 driver. In each category, each trip record is taken as the clustering object. That is, for a driver, if his behavior features are stable, all his trip records will be clustered into the same category. If there are changes in driver behavior, trip records will be clustered into different categories. Therefore, this paper defines a driver average classification accuracy metric. If each trip record of a driver can be effectively classified into the same category, it indicates the highest classification accuracy. The definition formula for average classification accuracy is shown in Equation (4), where n_i is the total number of trips made by driver i, C is the classification result, c_il is the number of trips of driver i in class l, and max{c_il} represents the number of trips of the driver' s records clustered into the class with the most trips. In an ideal state, all driving records of a single driver would be classified into one class, so max{c_il} = n_i. m is the total number of drivers. Finally, the mean classification accuracy is calculated for all drivers.

An example of driver clustering accuracy is shown in Table 1 . Taking Table 1 as an example, Driver 1 only drove Vehicle 1, and these three trips had small behavioral differences, all being classified into Class 1, so the classification accuracy for this driver is 100%. Driver 2 drove Vehicles 2 and 3, with small behavioral differences between the two vehicles, and all records of Driver 2 were classified into Class 2, with classification accuracy of 100%. Driver 3 drove Vehicles 4 and 5, with three records of driving Vehicle 4 classified into Class 1 and two records of driving Vehicle 5 classified into Class 2, resulting in classification accuracy of $3/5 \times 100\% = 60\%$. The final average classification accuracy for the three drivers is $(100\% + 100\% + 60\%) / 3 = 86.67\%$.

To compare the advantages and disadvantages of the above methods, this paper set the number of clusters to 3 and applied the K-means algorithm to cluster the partitioned "trip" data, obtaining the K-means algorithm results. According to the classification accuracy verification method proposed in this paper, the average accuracy calculated by the K-means algorithm is 75%, while the accuracy of the Blondel algorithm is 92.5% when the threshold is determined, showing a significant improvement compared to the K-means algorithm.

3 Conclusion

This paper proposes applying the Blondel algorithm, a network community detection algorithm, to multidimensional data clustering analysis. Through analysis of actual bus operation data, a specific analysis scenario was designed where a bus trip from start to end point is defined as one trip. A network was constructed based on the similarity between attribute vectors of each trip' s driving behavior.

The Blondel algorithm was used to detect communities in this network to obtain clustering results. Through the proposed accuracy measurement method, the results show that the network community detection algorithm achieves an accuracy of 92.5%, outperforming the K-means clustering algorithm' s 75%. Multidimensional data clustering analysis has broad application prospects in numerous fields.

This paper uses network community detection methods to cluster multidimensional data, but the clustering effect on higher-dimensional data requires further research. At the same time, clustering algorithms for multidimensional data can be considered and studied from more perspectives.

References

[1] Zimek A, Schubert E, Kriegel H P. A survey on unsupervised outlier detection in high-dimensional numerical data [J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2012, 5 (5): 363-387.
[2] Sarstedt M, Mooi E. A concise guide to market research [M]. 2nd ed. Berlin: Springer, 2014: 273-324.
[3] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods for the k-means clustering algorithm [J]. Expert Systems with Applications, 2013, 40(1): 200-210.
[4] Kriegel H P, Kröger P, Sander J, et al. Density-based clustering [J]. Data Mining & Knowledge Discovery, 2011, 1(3): 231-240.
[5] Tran T N, Drab K, Daszykowski M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters [J]. Chemometrics and Intelligent Laboratory Systems, 2013, 120(15): 92-96.
[6] Ding Hongxu, Wang Wanxin, Califano A. IterClust: a statistical framework for iterative clustering analysis [J]. Bioinformatics, 2018, 34 (16): 2865-2866.
[7] Zhang Xiaolu. Pythagorean fuzzy clustering analysis: a hierarchical clustering algorithm with the ratio index-based ranking methods [J]. International Journal of Intelligent Systems, 2018, 33(9): 1798-1822.
[8] 任卓明, 刘建国, 邵凤，等. 复杂网络中最小 k-核节点的传播能力分析 [J]. 物理学报, 2013, 62 (10): 108902. (Ren Zhuoming, Liu Jianguo, Shao Feng, et al. Analysis of the spreading influence of the nodes with minimum k-shell value in complex networks [J]. Acta Physica Sinica, 2013, 62 (10): 108902)
[9] 杨剑楠, 刘建国, 郭强. 基于层间相似性的时序网络节点重要性研究 [J]. 物理学报, 2018, 67(4): 48901-048901. (Yang Jiannan, Liu Jianguo, Guo Qiang. Node importance identification for temporal network based on inter-layer similarity [J]. Acta Physica Sinica, 2018, 67 (4): 48901-048901.)
[10] 胡小军, 郭强, 杨凯,等. 基于相对熵的多属性作者学术影响力排名研究 [J]. 电子科技大学学报, 2018, 47(2): 279-285. (Hu Xiaojun, Guo Qiang, Yang Kai, et al. Multi-attribute researcher academic influence ranking based on relative entropy [J]. Journal of University of Electronic Science and Technology of China, 2018, 47(2): 279-285.)
[11] Bloomfield N J, Knerr N, Encinas V F. A comparison of network and clustering methods to detect biogeographical regions [J]. Ecography, 2018, 41(1):

1-10.
[12] Girvan M, Newman M E. Community structure in social and biological networks [J]. Proceedings of The National Academy of Sciences, 2002, 99 (12): 7821-7826.
[13] Newman M E, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E, 2004, 69(2): 026113.
[14] Papadopoulos S, Kompatsiaris Y, Vakali A, et al. Community detection in social media [J]. Data Mining and Knowledge Discovery, 2012, 24 (3): 515-554.
[15] Xie Jierui, Kelley S, Szymanski B K. Overlapping community detection in networks: the state-of-the-art and comparative study [J]. ACM Computing Surveys, 2013, 45 (4): 43.
[16] Fortunato S, Hric D. Community detection in networks: a user guide [J]. Physics Reports, 2016, 659 (11): 1-44.
[17] 杨凯, 郭强, 刘晓露, 等. 基于多重特征向量的有向网络社团结构划分算法 [J]. 电子科技大学学报, 2016, 45(6): 1014-1019. (Yang Kai, Guo Qiang, Liu Xiaolu, et al. Detecting community structure in directed networks via multiple eigenvectors [J]. Journal of University of Electronic Science and Technology of China, 2016, 45(6): 1014-1019.)
[18] Yang Kai, Guo Qiang, Li Shengnan, et al. Evolution properties of the community members for dynamic networks [J]. Physics Letters A, 2017, 381(11): 970-975.
[19] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.
[20] Myers T A. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data [J]. Communication Methods & Measures, 2011, 5(4): 297-310.

Note: Figure translations are in progress. See original paper for figures.