International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 12 (2014), pp. 107-116
http://dx.doi.org/10.14257/ijmue.2014.9.12.10

Subspace Clustering Algorithm Based on Multi-rule Constraint

Huiping Li

Bao Tou Medical College
Inner Mongolia Bao Tou 014040, China
lhpccc@163.com

Abstract

For the fact that telecom data size is extremely huge and the management is much complicated, the paper proposes subspace clustering algorithm based on multi-rule constraint, to mine business knowledge information in a more efficient and accurate manner. By relying on K-means clustering algorithm, the method improves selection and mutation operation of genetic algorithms and thus corrects inappropriate choice of K-means initial clustering centers. Meanwhile, with the use of variable weighting strategy, data classification sparseness in the clustering is overcome. A fast and useful mining method is enabled for massive data. Results show its better performance in terms of computing efficiency, accuracy and ability.

Keywords: *data mining, telecommunications area, k-means, cloud computing*

1. Introduction

Another challenge accompanied with the information society is occurrence of tremendous data every hour and moment. Those data play a critical role in company operation and business decision. As statistics suggest, the bill volume of China Mobile every single day reaches 1 TB for a prefecture-level city. Google needs to process worldwide webpages every day for over 20PB. Those data, if used effectively, will bring about unlimited values to enterprises. If they're neglected or of disregard, they'll become "data tombs", which would add enterprise operational costs [1-5]. There will have difficulties when enormous data are being discovered:

When data volume is considerably huge, tele-data up to hundreds of T-scale memory size is commonly seen [6-7]. In this case, it will extremely hard to perform knowledge discovery through traditional data mining tools. More unexpected problems will occur if complicated application tasks and enterprise requirements are combined together. To be specific, in normal running cases, programs will abruptly terminate and resources collapse, which are both anticipated [8-9].

More sophisticated processing methods and systematic planning are required. It's necessary to consider in a global sense the integrative performance of clustering algorithms [10-11]. In the process of treatment, lots of practical issues should be settled like: if the knowledge classification is appropriate; if business knowledge is accurate; how to achieve positively the best rate of precision and acquire business knowledge satisfactory to users [12-13].

To sum from the above, data size is inversely related to the performance of methods. With size growing, the comprehensive performance of methods will degrade linearly. It's required to strengthen the effectiveness of data mining techniques by plenty of rule and constraint mechanisms. Hence for the size of tele-data, it uses the global optimal features of genetic

ISSN: 1975-0080 IJMUE

algorithms and precise subspace clustering characteristics of variable weighting scheme to modify traditional K-means clustering methods, for a proper data mining technique which has better overall performance [14-15].

2. Application and Analysis of the Method

The subspace clustering is composed of four parts. To make it clear, we'll introduce the design and implementation procedure of it by taking the case of dynamic classification of telecom product suppliers.

Retrieve historical behavior records of those providers to prepare source data of that dynamic categorization;

Date pre-treatment, carry out normal operation of data to form features for describing behavioral data and establish the matrix;

According specific knowledge discovery demand and data features, the dynamic classifying strategy is developed on the basis subspace clustering of variable weighting;

Compare dynamic and static classification results as to adjust telecom operational strategies.

Behavior characteristic characterization: Conventional classifying methods use qualitative index and inherent attributes as major evaluation basis, without regard of behavior's internal association and interaction between telecom enterprises and suppliers. As plentiful features change along with trade volume, errors are directly caused in the discovery of final knowledge.

Data matrix: As time axis defines, tele-data can be described to behavior attribute set, *i.e.*, basic information of articles in every single given time range. Each to-be-discovered target can be expressed to a value set of those attributes. The whole target set can be expressed to a N x M matrix, where N is total number of targets and M is totality of behavior attribute. Table I lists the behavior data matrix of one supplier.

Object data is divided into C_0 and C_1. The behavioral pattern of targets in both groups is exactly similar. Columns refer to behavior attribute of objects; 0 and 1 respectively means the trade behavior which doesn't happen and has occurred.

Table 1. Behavior Data Matrix

Attribute / Grouping	i_0	i_1	i_2	i_3	i_4	C
S_0	1	1	1	0	0	C_0
S_1	1	1		0	0	
S_2	0	1	1	0	0	
S_3	0	0	1	1	1	C_1
S_4	0	0	0	1	1	
S_5	0	0	0	1	1	

In real situation, the matrix is of tremendous size. When it has high dimensions, data get much sparser. So for effective data mining, it's necessary to adopt subspace clustering

analysis method based on variable weighting strategy.

Variable weighting K-means subspace clustering method can meet knowledge discovery demands by overcoming hi-dimension, huge amount and data sparseness. With modification of data variable weight, there will be importance and less importance for different objects, instead of mere deletion or conservation. Furthermore, the method retains data processing ability of traditional K-means mining, especially suitable for discovery of bulky data like telecom information. The application of data behavior matrix can describes effectively the attribute information of various telecom business types, rather fit for mining data in the telecom field.

Set Data matrix with $N \times M$ contains M variable, the N object is divided into K type, set $\Re_i = (\delta_{i,1}, \delta_{i,2}, ..., \delta_{i,m})$ the weight of variable in the $i-th$ class; set $\Re = (\Re_1, \Re_2, ..., \Re_K)$ the weight collection of all variables in all K classes.

The variable weighted subspace clustering method in the calculation process automatically adjusts the variable weight of each cluster value. In order to meet the data needs, to reduce the classification problem of sparse impact on the clustering results. Formulation is expressed as:

$$F(W, Z, \Re) = \sum_{l=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{M} w_{l,j} \delta_{l,j}^{\beta} [d(z_{l,j}, x_{j,i}) + \eta] \quad (1)$$

As mentioned earlier, K is the number of clusters, $[W_{i,j}]$ is an integer matrix of $K \times N$. The variable Z represents a K cluster center, Z is the $K \times M$ integer matrix elements. $d(z_{l,j}, x_{j,i})$ represents object j to cluster center l distance at $i-th$ characteristic

As mentioned before, in order to overcome the data sparseness problem, once in the weight calculation. Introduced a steady 0 constant ε. In subspace clustering, in order to facilitate the processing of feature weights, introduces a parameter η. When η is too large, the algorithm will K back to the standard K-means clustering algorithm. When η is too small, the whole sample space will focus on the sparse higher degrees of variables, while neglecting other important variables. Therefore, selecting parameter η take specific manner, in general, meeting the following conditions:

$$\eta = \frac{\sum_{j=1}^{N} \sum_{i=1}^{M} d(x_{j,i}, o_i)}{N \cdot M} \quad (2)$$

Where, o_i represents characteristics average of the whole sample space. For the application of mass data, complex, it can be sampled to determine the parameter η size. After discussed, for the business application of this case, it can be sampled 20% to obtain satisfactory clustering results.

It solves data processing problems between data objects and in the process of knowledge discovery, capable to analyze and treat data dynamic classification. In the context of data exchange, K-means methods have the feature of rapid treatment of abundantly huge data. Subspace clustering methods can take advantage of it to process in a more efficient way hi-dimension data and completely get rid of the ubiquitous knot of data sparsity.

3. Experiment Design and Discussion

3.1. Test Scenario and Data

The test environment here is on VC.NET application to perform synchronous test. Crank call information mining was chosen as primary data for business analysis and validation. Still, comparison was made among several characteristics like algorithm computing efficiency, to determine the comprehensive performance of the subspace clustering algorithm based on multi-rule constraint. Crank call information mining is the process of all suspected strange call numbers to be accurately searched from tremendous telecom communication data, which solved in essence the actual situation that customer service staffs have to dial back to confirm nuisances, reduced their workload, enhanced the precision ratio of recognizing such calls, intercepted intelligently nuisance calls and blocked fully strange call numbers in the premise of limited blacklist capacity.

Sampling information of telephone exchange data is summarized in Table 2. To facilitate the experiment and test, we sorted out three morning hours' exchange data on one working day from exchange information of some prefecture-level city, which takes up approx. 3% of total volume on the day.

Table 2. Abstract of Sampled Data

Summary data	Numerical value
Call number	3945190
Telephone number	893202
Relates to the city regions	16233
Phone number back to dial the number	7441

3.2. Analysis of Telecom Business Application

Based on week to estimate the life circle of strange calls. Together with the help of harassment frequency and instance time, we can calculate the priority of being held. The first 100 blacklist numbers from other provinces in the priority ranking list, blacklist numbers in all networks and the life cycle are directed to webmaster interception system for barring. A few characteristics of nuisance call behavior information are extracted and the related data warehouse classification recognition model is created. Then, data are pre-treated with the proposed method. Through calculation to judge accurately the suspicious strange call numbers, the rate can reach over 80%. The traditional classifying method uses mainly static attributes, negligence of return of such calls. Owing to behavioral pattern difference, features of those harassments are not carefully concerned. That's why we proposed the new method to analyze data relating to such call behavior, allocating phone numbers with the similar behavior features to one class:

(1) One side calling the other and holding time over 1.5mins;
(2) One side calling the other but ever answered;
(3) One side calling the other but hanging up instantly;

In return, we conclude these behavior features of the other side:
(4)Unanswered call but dial-back and talk time over 1.5mins;
(5)Unanswered nor call back;
(6)Unanswered but dial-back and hang-up after 30 seconds;
(7)Answered but immediate hang-up;
(8)Answered and conversation time over 1.5mins

According to the above classification, we can positively figure out suspicious strange call blacklist. The whole procedure is described as follows:

Step 1: Give M variables in $N \times M$ data matrix and partition call behavior of all numbers into K classes. Set $\Re_i = (\delta_{i,1}, \delta_{i,2}, ..., \delta_{i,m})$ the weight of variable in the $i-th$ class; set $\Re = (\Re_1, \Re_2, ..., \Re_K)$ the weight collection of all variables in all K classes. In the instance, M=24; N=893202. From call behavior of those numbers in different time, we can decide the total number of classes is 8. Set clustering parameter K=8.

Step 2: Dynamic behavior data are normalized as to establish data matrix, such as call behavior, in-coming time, prefectural city area and talk time of all phone numbers.

Step 3: Use the proposed algorithm based on multi-rule constraint to cluster the classifying method of phone numbers, which can be realized through classification model formula 1.

Step 4: According to formula 2, the value of η can be defined. Set feature weight. In this process, for the call behavior in the above viii, the weight of (3) and (7) is bigger.

Step 5: Based on call behavior similarity, all phone numbers are clustered into one class as per their similarity percentage, to form different clusters.

Clustering results are seen in Figure 1, where horizontal axis is classification K of phone numbers; vertical axis is quantity of phone numbers included in each class. The bar chart displays the status of all original static classification information after dynamic clustering results are analyzed.

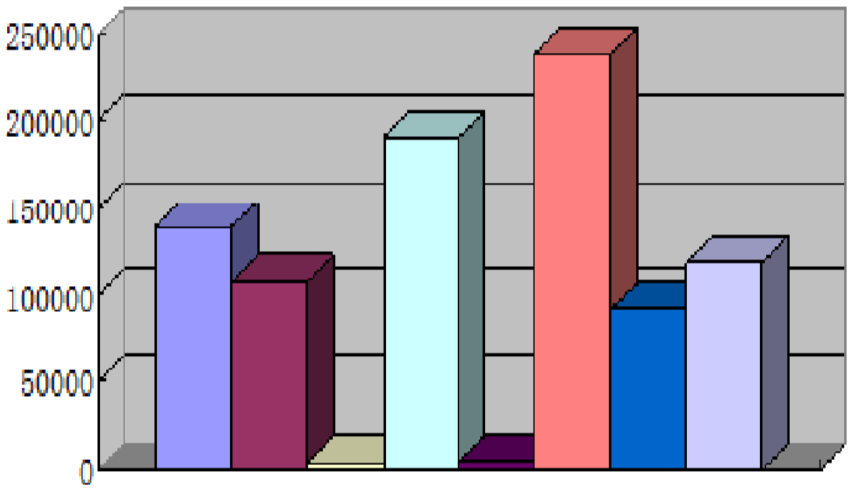

Figure 1. Results of classification algorithm

For phone numbers formed from (3) and (7), an effective discernment model can be set up and the blacklist is produced to be used in the nuisance phone number monitoring system. Therefore, it's necessity to change business operation decision in accordance to results of classification (Table 3).

Table 3. Countermeasures Classification

Cluster number	Countermeasures
1，4，8	maintain the original management strategy
2，5	Adopt the same management strategy
3，6，7	Further dividing the types, targeted management

Through strategy adjustment, the monitoring blacklist of harassment phone numbers can be formed and refined to improve the accuracy of judgment. As seen in Figure 2, the proposed method is employed to do knowledge discovery of uncertain strange phone numbers, with the accuracy ratio of that determination more than 90%.

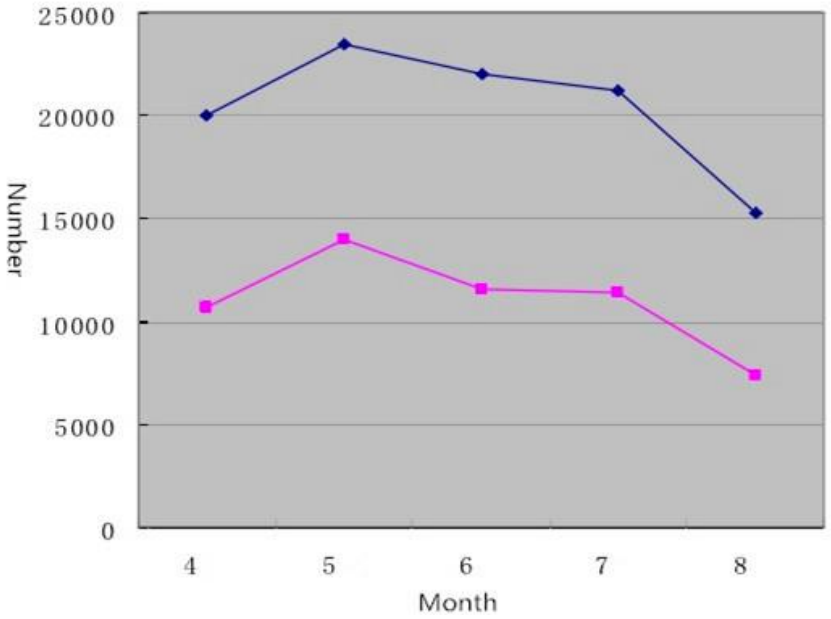

Figure 2. Statistical Number of Monitoring Harassing Phone Calls

In Figure 2, the quantity of harassment phone numbers becomes fewer and fewer since May, at 11.63% per month on average, time for discerning reduced by 963.5mins per capita. The subspace clustering technique based on multi-rule constraint, the detection ratio of strange calls is remarkably raised. The rate of complaints by call-back phones is hugely decreased.

3.3. Data Magnitude Test and Analysis

As observed from Table 4, to prove the accuracy of the proposed method here, we compared it with traditional K-means algorithm on clustering results. The test has two phases.

Stage 1: We chose methods which only have initial clustering center optimization strategy but variable weighting function to make comparison, shown in Table 4 (a);

Stage 2: We chose complete subspace clustering algorithm based on multi-rule constraint to make comparison, listed in Table4 (b).

What's examined by those algorithms is largely degree of difference. When two methods are for calculation, and one object is assigned to different clusters, the object is called differential sample. The ratio of differential samples in the total samples is called the degree of difference.

Table 4. (a) Results Comparison between Proposed Algorithm and K-means for 1st Phase

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
Total number of samples	910	870	620	1011	768	834	697	881	813	948	8352
Number of Differences samples	30	42	36	44	28	42	26	34	40	62	384
Difference degree (%)	3.3	4.8	5.8	4.3	3.6	5.0	3.7	3.8	4.9	6.5	4.6

Table 4. (b) Results Comparison between Proposed Algorithm and K-means for 2nd Phase

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Total
Total number of samples	910	870	620	1011	768	834	697	881	813	948	8352
Number of Differences samples	7	12	7	21	15	7	11	14	18	23	145
Difference degree (%)	0.8	1.4	2.7	2.0	1.9	0.8	1.5	1.5	2.2	2.4	1.7

3.4. Performance Validation and Analysis

Here we chose Data and Iris which are corresponding to simulation and benchmark dataset to confirm the integrated performance of the proposed method. Table 5 presents the size, dimension and cluster size of dataset Iris and Data.

Table 5. Descriptions of Data Sets

Name	Size	The number of levels	Dimension	Class size
Data	1000	4	4	250,250,250,250
Iris	1500	3	3	500,500,500

Figure 3 shows the distribution of data set. Dimensional characteristics of Iris dataset makes it impossible to show a chart form. The collection contains more than 150 models, each model contains 4 attributes.

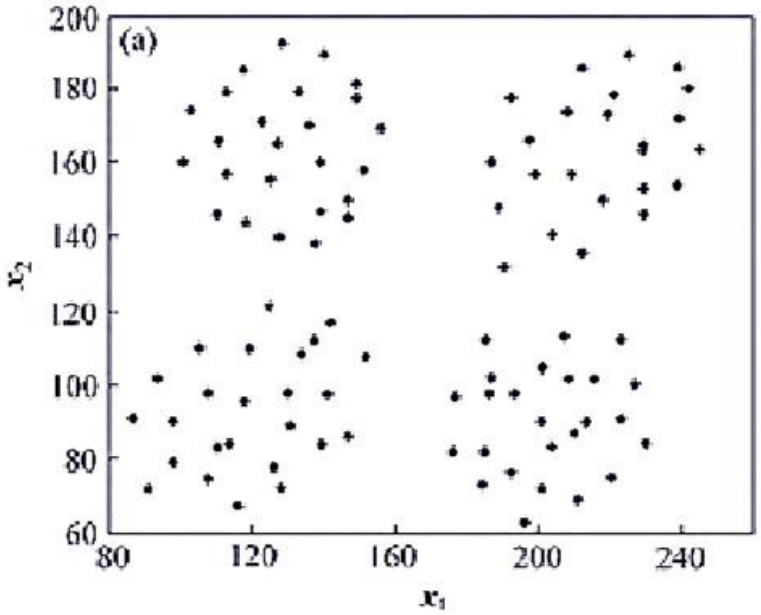

Figure 3. Distribution of Data Set

For the traditional K-means method, comparisons were made between the optimized method for the initial clustering center and subspace clustering method based on multi-rule constraint. Each was computed for three times. After comparisons, we found:

The proposed method increased apparently the total number of clustering classes;

The optimized method for the initial clustering center presents more clear structure of clustering classes than the traditional K-means method.

Table 6. Experimental Analysis for Three Kinds of Algorithms

Algorithm	Data set	Run time (s)	Accurate rate	Find the number of levels
K means algorithm	Data Iris	24.30-26.43	90.28%	4,3,3, 2,3,2
The initial cluster center. K means method	Data Iris	12.22-13.98	94.81%	4,4,4,3,3,3
Multi rule constraint subspace Clustering algorithm	Data Iris	9.02 -11.64	97.73%	4,4,4,3,3,3

Table6 is the test environment three kinds of clustering algorithm in the data set standards under the experimental analysis, Table 5 gives the average running time and accuracy.

3.5. Emulation Test

For the emulation, we selected traditional K-means algorithm, GA-based K-means method and subspace clustering method proposed here. The experiment used VC development in MATLAB environment. In the test, for the genetic algorithm, we set crossover probability 0.76, mutation possibility 0.08, group size 1000 and 100 iterations. When the number to obtain the best result is consecutively up to 10 times, it's over. The test dataset is still Iris, of 150 samples. Suppose every method of the three run 10 times to get the maximum, minimum

and mean values in the intra-cluster distance. On that basis, we calculated the maximum, minimum and mean values in the inter-cluster distance. Results are all seen in Table 7.

Table 7. Experimental Results for Iris Data Sets

Algorithm	The maximum distance cluster	The minimum distance cluster	The average distance of clusters	The maximum distance between clusters	The minimum distance between clusters	The average distance between clusters	Optimal iteration
K means algorithm	13.392	9.492	11.832	9.201	5.081	5.920	46
K means algorithm based on genetic algorithm	13.209	9.313	11.642	9.892	5.293	6.082	37
This paper algorithm	13.002	7.498	11.105	10.174	4.801	6.301	20

For clustering methods, the goal is to obtain both the minimum intra-cluster distance and the maximum inter-cluster distance. From Table 7, traditional K-means method depends more on the selection of initial center for the accuracy of clustering. Once it's selected wrong, it will fall into local optimum, leading to earlier convergence. For GA-based K-means method, too simplified data and worse diversity caused pre-mature, disable to cope with the problem of data sparseness. The proposed method keeps the search ability of genetic algorithm in global area and has the power to solve sparse data to acquire data of higher quality.

4. Conclusion

This paper analyses the defects and core problem in traditional K means methods. That is difficult to select the initial cluster centers and data processing classification sparse case. For these two problems, we proposed corresponding solutions: to achieve the global optimal search using the genetic algorithm, to avoid falling into local optimum clustering results. Improved subspace clustering strategy using the variable weighted, it can handle massive high dimensional sparse data. On this basis, the proposed method makes a practical analysis combined with telecom specific business, and compares it with the traditional clustering algorithm, the performance of the paper algorithm are verified the accuracy and rationality.

References

[1] Y. Xiong, "Analysis of the telecommunications industry management system construction discussion", The accounting &OSS world, no. 11, **(2004)**, pp. 43-44.
[2] B. Rongbing, "Research and application of grid density based on data mining", Master Thesis of Nanjing Institute of meteorology, **(2003)**, pp. 12-19.
[3] O. Grupe, "DATA BASE MINING", Discovering New Knowledge and Competitive Advantage, **(2005)**, pp. 22-24.
[4] C. Ping, W. Bo and X. Liutong, "Parallel algorithm and scheduling algorithm of grid Telecom social network intermediary degree", Journal of Beijing University of Posts and Telecommunications, no. 12, **(2006),** pp. 34-37.
[5] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large database", Proceeding of the 2003 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, **(2003)**, pp. 207-216.

[6] Y. Xiang, Y. GUI, X. Xiandong and W. Jianwei, "A flow of subspace clustering method based on the data of regional division", Journal of computer research and development, vol. 01, **(2014),** pp. 88-95.
[7] X. Yingmei, "Research on clustering algorithm for dynamic data flow within the sliding window", Journal of Shaanxi University of Technology (NATURAL SCIENCE EDITION), vol. 01, **(2014),** pp. 42-46.
[8] X. Y. Xi, H. Huan, J. Jin and Y. Huaiwang, "A kind of uncertain data stream subspace clustering algorithm", Information technology, vol. 02, **(2014)**, pp. 27-30.
[9] L. You, L. Wei, L. Junzhou, J. Jian and X. Anger, "A selective cooperative learning network user behavior anomaly detection method", based on Chinese Journal of computers, vol. 01, **(2014)**, pp. 28-40.
[10] W. Lijuan, H. Zhifeng, C. Ruichu and W. Wen, "Algorithm locally adaptive subspace clustering based on multi disturbances", Computer science, vol. 02, **(2014)**, pp. 240-244.
[11] L. Weisheng, G. Gongde and C. Lifei, "SMwKnn: mutual k nearest neighbor algorithm based on weighted subspace distance", Computer science, vol. 02, **(2014)**, pp. 166-169.
[12] L. Tao, W. Weiwei and Z. D. Jia, "Weighted sparse subspace clustering image segmentation method", Systems engineering and electronics, vol. 03, **(2014)**, pp. 580-585.
[13] W. Ailian, W. Weili and C. Junjie, "Comparison and improved image segmentation method based on K-means clustering algorithm", Journal of Taiyuan University of Technology, vol. 03, **(2014)**, pp. 372-375.
[14] W. Lijuan, H. Zhifeng, C. Ruichu and W. Wen, "A real value link analysis fusion algorithm based on ESSC", The research and application of computer, vol. 05, **(2014)**, pp. 1366-1369.
[15] S. Aihui, H. Shucheng and L. Sweet, "A clustering algorithm of network community structure and module function", Based on the computer application and software, vol. 04, **(2014)**, pp. 74-277.

Author

Li Huiping, she received her master's degree of engineering in Inner Mongolia Normal University. Now she is a lecturer in Bao Tou medical college. She is in the research of computer application.

