A Novel Clustering Algorithm Based on Graph Matching

Guoyuan Lin
School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
Email: lingy@cumt.edu.cn

Yuyu Bie and Guohui Wang
School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
Email: {bieyuyu@126.com, wangguohui00@sina.com}

Min Lei
Information Security Center, Beijing University of Posts and Telecommunications, Beijing, China
National Engineering Laboratory for Disaster Backup and Recovery, Beijing University of Posts and Telecommunications, Beijing, China
Email: leimin@bupt.edu.cn

***Abstract*—Aiming at improving current clustering algorithms for their failure to effectively represent high-dimensional data, this paper provides a novel clustering algorithm-GMC-based on graph matching with data objects being represented as the attributed relational graph and the graph matching degree being the standard of similarity measurement. In the algorithm, graphs for classification will be matched with character pattern atlas, and classified into the class with the biggest similarity. The accuracy and rationality of this algorithm is always kept with continuous renewal of character pattern atlas. In addition, compared with the classical K-means clustering algorithm and Newman fast algorithm, this algorithm shows its own superiority and feasibility in applications of data mining.**

***Index Terms*—clustering analysis, association rules, attributed relational graph, similarity matching, character pattern graph**

I. INTRODUCTION

Clustering analysis [1] is one of the most common technologies in the area of data mining. Clustering analysis involves many areas such as geometry, statistics, machine learning, KDD, etc. Clustering analysis can classify and describe data objects and their relations by information found in the data itself. The main purpose of clustering is data classification. Similar data objects will be assigned to the same group, while different data objects will be assigned to different groups [2].

The data set in the application database shows some high-dimensional data characteristics, such as complex property values, large data volume and sparse value distribution. High-dimensional data clustering has become a very important research topic in data mining. Authors in reference [3] have proposed a parallel clustering algorithm, named Stem-Leaf-Point Plot Clustering Algorithm (SLPPCA). Considering some characteristics of high-dimensional database, SLPPCA tend to produce clusters of different shapes and sizes exploiting the data-parallelism of data objects and adopting a task decomposition design to balance the workloads of multi-core processors.

The existing clustering algorithms cannot directly cluster the high-dimensional data, but with the data object being viewed as a point in the multidimensional space, and then based on the use of a linear function, geometry and statistics for data dimensional reduction. When dealing with high-dimensional data, dimension reduction will inevitably cause some problems. For example, the time complexity will increase with the growth of data dimensions; the dimension reduction will result in the loss of similar information; meanwhile the distance function will be difficult to define.

Therefore, in order to solve these problems causing by high-dimensional data, this paper intends to use graph matching technology for clustering analysis. Graph matching technology is one of the central topics in graph theory as well as in the theory of algorithms and their applications. The goal of graph matching is to determine whether two graphs are similar or not, while data classification of clustering analysis is on the basis of similarity matching. Besides, graph matching algorithm has major advantages when dealing with high-dimensional data objects. Therefore, graph matching can be well applied in the clustering area. Based on graph matching, a novel clustering algorithm (GMC) with a new data object representation model and similarity measure is presented in this paper. This clustering method based on graph matching not only can solve problems caused by high-dimensional data objects,

doi:10.4304/jsw.8.4.1035-1041

but also can improve accuracy and decrease the algorithm complexity.

The rest of this paper is organized as follows. In section II several researches of clustering technology especially graph clustering will be discussed and analyzed. And then with the general idea of GMC algorithm described in section III, section IV presents a novel clustering algorithm based on graph matching. Section V explains the experiment results and analysis. Finally, the conclusions are summarized in section VI.

II. Research of Clustering Technology

The Internet is offering us huge amount of information, and in the meantime it throwing us to the challenge of better classification technology of these large-scale information. Clustering technology emerged as the times require. Mathematical method would be used in clustering technology to classify all kinds of data objects based on their characteristics and similarities.

The clustering technology is witnessing great evolution. The relationships and similarities of data objects of traditional clustering technology are usually not definite in reality. Aiming at the flaws of dynamic clustering, Jingtao Sun [4] presented a fuzzy clustering algorithm based on factor analysis method, which combines the technology of reducing dimension. This fuzzy clustering method can resolve practical problems and is applicable to E-mail filtering. However, it is not good enough to deal with high-dimensional data objects.

With the rapid development of network technology, the scale of graph which used to describe real-world network structure shows exponential growth. This phenomenon inspires many researchers to try to apply figure knowledge into the clustering technology.

In recent years, many scientists provide some effective methods based on similarity measure of the graph theory and clustering algorithms. In 2004, Vempala introduced the graph clustering information [5], and Kernighan-Lin algorithm based on the graph segmentation was proposed, which randomly divided notes in the graph into two sub-graphs of known size, and introduced a definition of gain function that the internal edges of sub-graphs minus edges connecting two sub-graphs, then constantly looked for division methods to maximize the gain function value. Hartuv and Shamir proposed a recursive method of cutting edge based on the graph theory [6], which was on the basis of the connection similarity of each node, by continuously removing edges from the graph, and divided the graph into various sub-graphs; Newman put forward GN algorithm and Newman fast algorithm based on the edge betweenness reduction [7], and the basic idea of which was that for each graph category, calculating the betweenness of each edge and removing the edge with the largest betweenness, and then calculating the betweenness of the remaining edges, and deleting the edge with the largest betweenness until all edges had been removed.

Graph clustering technology is a promising study field. However, it is confronted with severe problems and challenges.

(a) Current representation methods of data objects in graph clustering cannot effectively represent the similarity relationship of each attribute;
(b) The similarity measurement of graph nodes is still based on the distance function, and the clustering result is inaccurate;
(c) The choice of the initial sub-graph influences clustering effect.

In order to avoid the above shortcomings, this paper based on graph theory innovatively represents data objects as the attribute adjacent graph, and the graph matching degree as the standard of similarity measurement. This method can not only effectively solve problems in the multidimensional clustering but also ensure the accuracy of clustering results.

III. General idea of GMC algorithm

Case-based Reasoning (CBR) mechanism is an artificial intelligence technology proposed in the early 1980s, which extracts similar cases from cases, and obtains solutions by the reference of similar cases [8].

CBR consists of three steps:
(a) Case library construction;
(b) Similar case search;
(c) CBR reasoning.

Our graph clustering algorithm proposed in this paper refers to CBR reasoning mechanism, including the study of data objects representation, similarity measurement, matching principle, characteristic mode atlas, etc.

Similarity measurement between data objects is the basis of clustering analysis, which directly affects clustering results. It is very difficult to have similarity identification for the high-dimensional data, therefore, the algorithm proposed in this paper represents data objects as the attributed relational graph , the greater the similarity graph matching is, the greater the similarity of data objects is, and vice versa.

The general idea of GMC algorithm is as follows:
(a) Initializing the attributed relational graph;
(b) Representing data for classification as the attributed relational graph;
(c) Graph matching with characteristic mode atlas G by matching principle:
  (i) If the graph matching succeed, which means that similarity data is found, and add it to the relevant class; Then update character pattern atlas G by the case library maintenance mechanism;
  (ii) If the graph matching fail, this graph is seen as a new class joining the training set; then update character pattern atlas G by the case library maintenance mechanism.

The general idea of GMC algorithm is shown in Fig.1. The case maintenance mechanism mainly refers to finding class center again and updating note matching code in characteristic mode figure and so on.

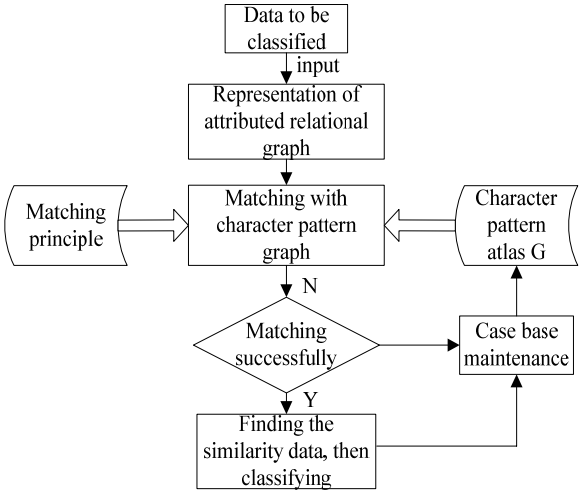

Figure 1. The general idea of GMC algorithm

IV. GMC CLUSTERING ALGORITHM BASED ON GRAPH MATCHING

A. *Data Objects Represented as Attributed Relational Graph*

Aiming at improving traditional clustering algorithms for their failure to effectively represent high-dimensional data, in this paper data objects are represented as attributed relational graphs, based on the feature recognition method.

The process is as follows:

Using one-dimensional segmentation technology to split n-dimensional data objects $\{A_1, A_2, \cdots, A_n\}$ to those objects with a single attribute value $\{(A_1,0,\cdots,0),(0,A_2,\cdots,0),\cdots,(0,0,\cdots,A_n)\}$, eac h object with a single dimensional value as a note, connects with each other and forms an attributed relational graph with n edges, then a data object can be described as $G(V,E,W)$.

In this process, V is the set of single attribute object V_i; E is the set of the edge e_i between node V_i and V_j; $W(e_i)$ is the attribute matching relation code which is defined in Segment B of peak V_i.

One-dimensional segmentation will bring a lot of problems in data storage. However, the relational graph constructed in this paper has its own advantages of sparser edges, more edge information and storage saving by using the relational table as a storage structure.

B. *Matching Principle*

GMC algorithm adopts graph matching similarity as the matching basis. The greater the similarity is, the more graphs matches. First, the attribute range is divided into different intervals. It is the offset degree of the corresponding note in the same range interval that is calculated by the matching relation code and thus to determine the matching similarity between graphs. The smaller the offset degree is, the larger the matching similarity is.

a. Matching relation code and the division of the range interval

Matching relation code is a binary number. For discrete attributes, if attribute values of the node V_i is the same as that of the same note in the matching graph, then the matching relation code is 1 and 0 otherwise. For continuous attributes, if the attribute values of the node V_i are in the attribute range of the same note in the matching graph, the matching relation code is 1 and 0 otherwise.

The value of the matching relation code depends on the division of the range interval. High-dimensional data contains discrete attribute ones as well as continuous attribute ones. If the attribute range is different, so is the division method. For discrete attributes, the attribute value is always turned to multiple numerical values from 0 to n (0 to n represent different attribute characters), and the value of range interval is the value of each numerical attribute; for continuous attributes, the algorithm in this paper adopts data association rules to extract principles in the process of its discretization, and then the attribute range is divided into intervals with a certain method.

If $I=\{i_1,i_2,\cdots,i_n\}$ is the set of data attributes, $D=\{d_1,d_2,\cdots,d_n\}$ is the set of data, and d_i is a data item in the set, then data association rule is as follows: $X \to Y$, among which, $Y \subset I$, and $X \cap Y = \Phi$. Association rules set two attributes [9]: support and confidence.

Definition 1 (support): the proportion of data items containing $X \cup Y$ accounting for all data items is denoted by:

$$s = p(X \cup Y) = \frac{|X \cup Y|}{|D|} \tag{1}$$

Definition 2 (confidence): the proportion of data items containing Y accounting for data items containing X is denoted by:

$$c = p(Y \mid X) = \frac{|X \cup Y|}{|X|} \tag{2}$$

How to set the support and confidence threshold will affect the accuracy and validity of classification methods. If the value is set too high, then the range intervals of sparse character items cannot be divided effectively; if it is set too low, then the range intervals of frequent character items cannot be divided effectively.

To meet data association rules, Fukuda put forward a method to divide attribute range into intervals with all the same depth by using the continuous attribute discretization principle [11]. This method by setting

effective support and confidence can decide the number of continuous attribute range intervals. However, each range interval of continuous attributes will be different, so is the size of divided intervals.

b. Offset degree

1) The definition of graph nodes offset degree is

$$OffV_i(G_p,G_q)=\frac{Sum\{W(e_i)_{G_p}\wedge W(e_i)_{G_q}\}}{\sum_{j=1}^{n}Sum\{W(e_j)_{G_p}\wedge W(e_j)_{G_q}\}} \quad (3)$$

In the equation above, $OffV_i(G_p,G_q)$ is the offset degree of graphs p and q on notes, and the range is [0, 1]. The matching degree is 0, which means that two nodes are not similar in the range (0, 1]. Therefore the smaller the offset degree is, the larger the matching degree is; $W(e_i)_{G_p}$ is the matching code of the i_{th} note in graph q and the character pattern graph; $W(e_i)_{G_p}\wedge W(e_i)_{G_q}$ is the collation operation of two matching codes; the function $Sum(\ )$ sums up the number of binary number 1 in the bracket.

2) $OffG_{pq}$ is the offset degree of Graph q and p and is defined as follows:

$$OffG_{pq}=\frac{\sum_{i=1}^{n}w(e_i)_{G_p}OffV_i(G_p,G_q)}{\sum_{i=1}^{n}w(e_i)_{G_p}} \quad (4)$$

The smaller the value is, the larger the graph matching similarity degree is.

3) Ave is the mean value of offset degree of graph p and character pattern atlas set and is defined as follows:

$$Ave=\frac{\sum_{q=1}^{k}OffG_{pq}}{k} \quad (5)$$

Here k is the number of graphs in the character pattern graph atlas.

C. The Construction of Character Pattern Graph Atlas

Character pattern Atlas $G=(G_1,G_2,\cdots,G_n)$ is obtained by handling the training data set. Given training data class set $D=(D_1,D_2,\cdots,D_n)$, each data record has been assigned to a particular class of D_i, the process is as follows:

1) Getting class centers

The data object potential can to a certain extent reflect its geometry features in the character space. According to this, Chiu proposed a method using potential function for class centers [13].

For a class with n data objects $D_i=\{d_1,d_2,\cdots,d_n\}$, among which, the potential function of d_i point is:

$$P_i=\sum_{j=1}^{n}e^{-\alpha\|d_i-d_j\|^2} \quad (6)$$

In $\alpha=4/r_\alpha^2$, r_α is a constant, and the radius of neighborhood. Data out of the neighborhood radius have little effect on the calculation of this point potential. The more points gather around data points, the higher its potential is. The highest point of the potential is the class center.

Chiu took r_α as a constant. Studies have shown that, due to the irregularity of the sample space, the choice of r_α should relate to the distribution of sample collection. There are usually two forms of neighborhood radiuses:

$$r_\alpha=\frac{1}{2}\min\{\max\{\|x_i-x_j\|,i=1,\cdots,n\},j=1,\cdots,n\} \quad (7)$$

$$r_\alpha'=\frac{1}{2}\sqrt{\frac{1}{n(n+1)}\sum_{j=1}^{n}\sum_{i=1}^{n}\|x_i-x_j\|^2} \quad (8)$$

In the equations above, r_α is half the distance from the middle sample in the sample collection to the sample farthest away from the center. r_α' is half the distance of the mean square root. Potential function is on the basis of exponentiation. When dealing with large data sets, its computing space is not fast. In order to avoid this, this paper takes r_α as the neighborhood radius.

2) Getting the character pattern graph

The training data set $D=(D_1,D_2,\cdots,D_n)$ is represented with attributed relational atlas $G'=(G_1',G_2',\cdots,G_n')$.

Here G_i' is the classified Atlas. When updating matching relation codes of all graph notes in G_i' and corresponding notes in the note center, if their attribute value is the same or in the same range interval, the matching relation code is 1 and 0 otherwise. Once updating all matching relation codes of classified graph notes in G', character pattern atlas $G=(G_1,G_2,\cdots,G_n)$ can be obtained.

D. Specific Steps of GMC Algorithm

GMC algorithm is an approach, in which data objects are represented as attributed relational graph and matching degree is taken as the standard of data similarity.

Specific steps are as follows:

Step1. Setting support and confidence thresholds, and constructing character pattern atlas G;

Step2. Inputting data for classification represented with attributed relational graph. The initial matching relation code of each note is set to 0;

Step3. $k=1$;

Step4. Taking the class center of the k-th class in character pattern atlas, and updating the matching node of each note in the graph for matching;

Step5. Matching all character pattern graphs in the k-th class with graphs for matching; obtaining the mean value of matching degree and saving it into $s[k]$;

Step6. $k++$;

Step7. If $k<n$ (n is the number of classes in the training set), then turn to Step3;

Step8. If $s[i]=0\ (0<i<k)$, then take this graph as a new class, and turn to Step8;

Step9. Comparing values in $\{s[1], s[2], \cdots, s[k]\}$, and classifying graphs into the class which has the smallest offset degree mean;

Step10. Calculating the class center, and updating matching relation codes of all graph notes to make sure the reasonability of character pattern atlas, and turn to Step2.

V. Simulation Results and Analysis

A. Experimental Data

For the evaluation of the clustering effect of GMC algorithm, experiments choose KDDcup99 data set which is often used in data mining. "kdd-train-nor" packet in KDDcup99 data set is taken as a training data set. Afterwards sampling 10% of the data set and getting "kddcup.Data_10_percent" packet for clustering. Experimental platform settings: Intel Core 2, memory 2G, Windows XP (sp3).

Each data in KDDcup99 data set contains 42 character attributes, the first 41 of which are fixed attributes of data object, and the last one is the class mark. The mark represents which class this data object belongs to.

B. The Settings of Support and Confidence Thresholds

The division of attribute range intervals is the key to graph matching. The settings of support and confidence thresholds determine the accuracy of division methods. In order to overcome the negative impact on final classification results due to inappropriate division of attribute range intervals, the support and confidence thresholds of the algorithm choose two groups of values with higher accuracy in the reference [12], namely, s = 0.2, c = 0.8 and s = 0.19, c = 0.76.

C. Results Analysis and Performance Evaluation of the Algorithm

Experiments adopt Matlab6.5 language, and use GMC algorithm, K-means algorithm, Newman fast algorithm for clustering data sets respectively.

Table I shows the comparison of clustering results with different test data volumes:

From Table I, when the data volume is 10 million, the clustering accuracy of GMC algorithm is lower than 74.1% of K-means algorithm and 73.7% of Newman fast algorithm. As K-means algorithm and Newman fast algorithms input less initial class numbers, while GMC algorithm has formalized representation of data objects and the design to continuously update the training set, these make the clustering accuracy of GMC algorithm be constantly improved with the increase of data volume to 89.7% and 85.6%, which are higher than 81.5% of K-means algorithm and 83.3% of Newman fast algorithm. With different data volumes, the clustering accuracy when s = 0.2 and c = 0.8 is obviously higher than that when s = 0.19 and c = 0.76.

TABLE I.

Algorithm Clustering Results Comparison

Data volume (million)	Clustering accuracy (%)			
	GMC algorithm (s=0.2, c=0.8)	GMC algorithm (s=0.19, c=0.76)	K-means algorithm	Newman algorithm
10	73.5	73.1	74.1	73.7
20	75.4	75.2	75.3	74.9
30	82.3	77.9	77.2	76.4
40	85.6	79.3	78.8	78.5
49	89.7	85.6	81.5	83.3

The algorithm performance was evaluated by Boley's clustering entropy [13], that is, the smaller the value of clustering entropy is, and the better the clustering effect is.

Suppose the data set G is divided into n number classes $G=\{G_1, G_2, \cdots, G_n\}$, in which $n|G_i|$ means data numbers in the class G_i, and $n|T_i, D_i|$ means data numbers of data attribute value T_i in the class G_i.

For the class G_i, clustering entropy $e(G_i)$ is defined as:

$$e(G_i)=-\sum_j \frac{n|T_i, G_i|}{n|G_i|}\log\frac{n|T_i, G_i|}{n|G_i|} \tag{9}$$

The overall cluster entropy is defined as a weighted average of all cluster entropies:

$$e=\frac{1}{\sum_{i=1}^{m} n|G_i|}\sum_{i=1}^{m} n|G_i|\,e(G_i) \tag{10}$$

When taking support and confidence threshold values in two different cases, the difference of GMC algorithm performances is shown in Fig.2.

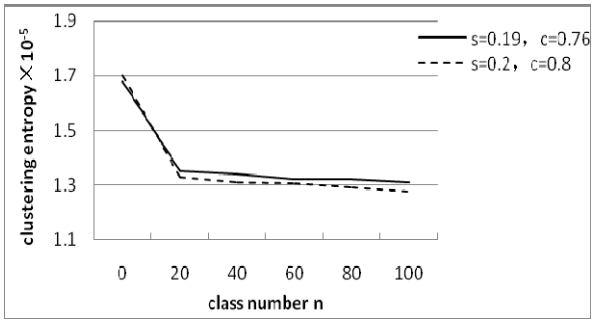

Figure 2. The difference of GMC algorithm performances

As shown in Fig.2, when the class number n is less than 30, the algorithm with two different threshold settings converge rapidly; when n is more than 30, the convergence is relatively stable. Besides, when the support threshold is 0.2 and confidence threshold is 0.8, the entropy value of the algorithm is lower than that with another setting, especially when n is more than 70, the entropy value of the clustering algorithm begins to be less than 1.3×10^{-5} which shows a better algorithm performance.

The support and confidence thresholds are 0.2 and 0.8 respectively, the performance comparison of GMC algorithm with the K-means algorithm, Newman fast algorithm is shown in Fig.3.

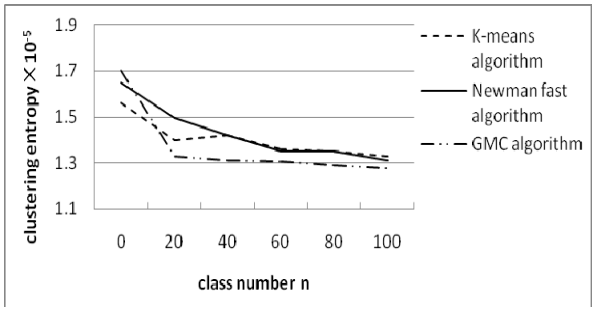

Figure 3. The algorithm performance comparison

As can be seen from Fig.3, when the class number n is less than 25, the clustering entropy of GMC algorithm reduces quickly from1.7×10^{-5} to 1.45×10^{-5}, showing a faster convergence rate; when n is more than 25, the clustering entropy of GMC algorithm is lower than those of K-means algorithm and Newman fast algorithm. Meanwhile, the general clustering entropy of GMC algorithm is significantly lower than those of others, and it has a better clustering result.

VI. CONCLUSIONS

This paper applying the graph knowledge into clustering algorithms defines a new method with a new data object representation model and similarity measure, and has proposed GMC clustering algorithm based on graph matching. This algorithm can not only directly cluster high-dimensional data, but also can gain a more accurate classification model based on the similarity measure of graph matching degree. Simulation experiments of KDDcup99 data set show that when dealing with high-dimensional data, GMC has a better convergence effect, and it is an effective method. In addition, the comparison of the clustering effect of GMC with those of K-means algorithm and Newman fast algorithm indicates that GMC algorithm has its own superiority and feasibility in applications of data mining.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their useful comments and suggestions on how to improve this paper. This paper is supported by the Opening Project of State Key Laboratory for Novel Software Technology of Nanjing University, China (Grant No.KFKT2012B25), and Youth Foundation by Beijing University of Posts and Telecommunications (Grant No.2013RC0308).

REFERENCES

[1] Min Xiao, Jijun Han, Debao Xiao, Zheng Wu, Hui Xu, "The overview of intrusion detection based on clustering," *Journal of Computer Applications*, vol. 28, no. 6, pp. 34-38, 2008.

[2] Jiawei Han, Micheline Kamber, "Data mining: concepts and techniques," *Morgan kaufamnn, San Francisco*, Aug. 2000.

[3] Jianfeng Yang, Puliu Yan, Yinbo Xie, Qing Geng, "An efficient parallel clustering algorithm for large scale database," *Journal of Software*, vol. 4, no. 10, pp. 1119-1126, Dec. 2009.

[4] Jingtao Sun, Qiuyu Zhang, Zhanting Yuan, "Fuzzy clustering algorithm based on factor analysis and its application to mail filtering," *Journal of Software*, vol. 4, no. 1, pp. 58-64, Feb. 2009.

[5] Ravi kannan, Santosh Vempala, Adrian Vetta, "On clusterings: good, bad and spectral," *ACM*, vol. 51, no. 3, pp. 497-515, 2004.

[6] Hartuv E. and R. Shamir, "A clustering algorithm based on group connectivity," *Information Processing Letters*, vol. 76, nos. 4-6, pp. 175-181, 2000.

[7] Newman M E J. "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, pp. 5-10, 2004.

[8] Barletta R. "An introduction to case-based reasoning," *AI Expert*, vol. 6, no. 8, pp. 43-49, 1991.

[9] Shenmiao Yuan, Xiaoqing Cheng, "The research of clustering method in data association rules finding," *Chinese Journal of Computers*, vol. 23, no. 8, pp. 866-871, 2000.

[10] Chiu S L. "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol. 2, pp. 267-278, 1994.

[11] Fukuda T, Morimoto Y, Morishita S, "Mining optimized association rules for numeric attributes," *In: Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. Montreal, Canada*, pp. 182-192, 1996.

[12] Jiao Zhang, Wei Wang, "The associative classification algorithm based on support and confidence threshold optimization technique," *Journal of Computer Applications*, vol. 27, no. 12, pp. 3032-3034, 2007.
[13] Boley D. L. "Principal direction divisive partitioning," *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 325-344, 1998.
[14] Jianwen Tao, "RCHIG: an effective clustering algorithm with ranking," *Journal of Software*, vol. 4, no. 4, pp. 382-389, Jun. 2009.
[15] Xiaoyong Liu, Hui Fu, "An effective clustering algorithm with ant colony," *Journal of Computers*, vol. 5, no. 4, pp. 598-605, Apr. 2010.
[16] Ulrik Brandes, Marco Gaertler, Dorothea Wagner, "Experiments on graph clustering algorithms," *Lecture Notes in Computer Science*, vol. 2832, pp. 568-579, 2003.
[17] Gary William Flake, Robert E. Tarjan, Kostas Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics* vol. 1, no. 4, pp 385-408, 2004.
[18] Taoying Li, Yan Chen, "Fuzzy clustering ensemble with selection of number of clusters," *Journal of Computers*, vol. 5, no. 7, pp. 1112-1119, Jul. 2010.
[19] Inderjit Dhillon, Yuqiang Guan, Brian Kulis, "A fast kernel-based multilevel algorithm for graph clustering," *KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM New York*, pp. 629-634, 2005.

Guoyuan Lin was born in Shandong, China in the year of 1975. He is now an associate professor at China University of Mining and Technology. He obtained his Ph.D. from Nanjing University in 2011. His research interests are information security and intrusion detection.

He works in China University of Mining and Technology for nearly 15 years since 1997. With the first author or instruction graduate student announce thesis more than 20 articles, among them abroad magazine's announcing to combine be registered a thesis 1 by the SCI and 11 by the EI. He published three books including: Computer Operating System (Beijing, China: Tsinghua University Press, 2011). His current research interests are information and network security.

Yuyu Bie is now a postgraduate at China University of Mining and Technology, Xuzhou, China. She was born in 1990, Shandong, China. She received the B.Eng. in School of Computer Science and Technology at China University of Mining and Technology in 2011. Her main research interests include information system security, access control and cloud computing.

Guohui Wang was born in 1987 in Heilongjiang, China. He obtained his Master's Degree in School of Computer Science and Technology at China University of Mining and Technology in 2012. His research interest is network security.