Journal of Theoretical and Applied Information Technology
20th March 2013. Vol. 49 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RESEARCH ON MULTI-RELATIONAL CLUSTERING PROBLEMS BASED ON PROPER-LINK

[1]YAN YAN, [2]AIMIN YANG, [3]YUANYUAN CAI

[1] Lecturer., College of Sciences ,Hebei United University, Tangshan 063009, Hebei, China
[2]Assoc. Prof., College of Sciences ,Hebei United University, Tangshan 063009, Hebei, China
[3] Teaching Assistant.,Qing Gong College ,Hebei United University, Tangshan 063009, Hebei, China
E-mail: [1]yanjxky@126.com, [2]43698059@qq.com

ABSTRACT

A relational database contains a wealth of information, and multi-relational clustering can be obtained by adopting proper-link. The similarities between prospers and linked calculation target topples can be worked out through searching for relevant multi-relational prospers and links of user clustering target, then the clustering can be completed by selecting CLARANS. In order to promote the efficiency of multi-relational data clustering algorithm, two-stage multi-relational data clustering algorithm is proposed. The method of fuzzy clustering is used to divide the area by the similar feature of load increasing. The new division is promising to improve the result of evident degree of clustering index to power load, the weighted demarcating method is inducted.

Keywords: *Data Mining, Multi-relational Clustering, Proper-link Clustering*

1. INTRODUCTION

Multi-relational data mining, a research filed appeared in 1998, explores relational data representation, uncertainty reasoning algorithms and learning algorithms to solve complex problems of the real world in all areas with the technology of inductive logic program design, relational database, machine learning, Web mining and so on. [1]

With the development of information technology, data are generated all the time. Therefore, all walks of life will accumulate large amounts of data. Through the use of data mining technology, some knowledge, which is interested in by some people, implicit, previously unknown, and with potential value in decision-making, can be extracted from these data. It is always an important sub-domain of artificial intelligence research to study the data by means of a computer program to get the implied message.

Multi-relational data mining methods can be divided into two categories: one is to convert the multi-relational data into single-relational data, and then applies the traditional data mining methods. There are two ways to convert multi-relational data into single-relational (single-table) data: one is to create a full name relation, and all the data are added into one table to form single-relational data; another is to create a new attribute in the center relation to form single-table. The advantage of this method lies in that the existing data mining methods can be applied directly. However, during the process of converting multi-relation to single-relation, some problems may exist, such as the increasing size of data, the missing of potential data, the problem of data replication.

2. BASED ON TIME SERIES SIMILARITY RESEARCH IN DATA MINING

There is a wealth of information in a relational database, including the information of data itself and the information of relationships between data objects. Therefore, it is a subject worthy of study to find out how to make use of this information to complete the clustering mining. [2]

Definition 2.1 The similarity of two target tuples in attribute f is recorded as $Sim_f(t_i, t_j)$, then for the categorical attribute

$$Sim_f(t_i, t_j) = \sum_{k=1}^{l} f(t_i) p_k f(t_j) p_k$$

and numeric attribute, σ_h is the standard deviation of target tuple in attribute h, thus the similarity of the target tuples is

$$Sim_h(t_i, t_j) = \begin{cases} 1 - \dfrac{|h(t_i) - h(t_j)|}{\sigma_h}, |h(t_i) - h(t_j)| < \sigma_h \\ 0, |h(t_i) - h(t_j)| \geq \sigma_h \end{cases}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1. Similarity of Multi-relational Attribute

Attribute selection in the conventional pretreatment process is used to remove those redundant or inconsistent attributes. Here pretreatment refers to search the most relevant attributes and links of user clustering target. In other words, the similarity comparison method between multi-relational attributes should be given first, and then according to the results of comparisons, search algorithm between the multi-relational attributes and links is achieved, thus, the process is completed.

According to definition 2.1, the similarity calculation[3] method of two target tuples in any multi-relational attribute can be worked out. The similarity between tuples can be calculated by attributes, while, it can also consider a certain attribute similarity reflects its impact on tuple similarity, which is the degree of correlation between attribute and clustering target. Based on this consideration, the similarity between attributes can be worked out through target tuples. If there are N target topples in T_{target}, then for any attribute f, according to definition 2.1, the similarity $Sim_f(t_i,t_j)$ between any target tuples can be obtained. Therefore, attribute f can be represented by a N^2 dimensional vector V^f. So that N^2 dimensional vector can represent any attribute.

Therefore, the similarity between any attributes and can be defined as their vector cosine:

$$Sim(f,g)=\frac{v^f v^g}{\left|v^f\right|\left|v^g\right|},(\left|v^f\right|=\sqrt{v^f v^f}).$$

2.2. Optimized Similarity of Multi-relational Attributes

The v^f and v^g should be obtained firstly to work out the similarity of multi-relational attributes, which will cost a lot. In order to improve the calculation result of similarity, the algorithmic methods of different categories of multi-relational attributes are studied respectively, thus it is no need to generate v^f and v^g.

For the vector of categorical attributes, the problem of similarity of N^2 dimensional vector can be converted into the similarity between attribute values.[4] Similarity of target tuples is related to its corresponding attribute value, so the similarity between attribute values can be worked out by tuples. Thus, similarity between any categorical valu v_k of categorical attribute f and any categorical value v_q of categorical attribute g is

$$Sim(fv_k,gv_q)=\sum_{i=1}^{N} f(t_i)p_k g(t_i)p_q$$

Therefore, for $v^f \otimes v^g$, there is

$$v^f v^g=\sum_{i=1}^{N}\sum_{j=1}^{N} Sim_f(t_i,t_j)Sim_g(t_i,t_j)$$
$$=\sum_{k=1}^{l}\sum_{q=1}^{m} Sim(fv_k,gv_q)^2.$$

It only needs to calculate the similarity between each value v_k of f and v_q of g to work out the similarity between f and g. For any tuples, it can be got $f(t)$ and $g(t)$. If $f(t)v_k>0, g(t)v_q>0$, adding $f(t)v_k \otimes g(t)v_q$ continuously in the process of scanning target tuples, then $Sim(fv_k,gv_q)$ can be obtained. In this way, through one scanning process, $Sim(fv_k,gv_q),(1\le k\le l,1\le q\le m)$ can be worked out. However, in practice, categorical attribute value is limited, so $v^f \otimes v^g$ can be finished within linear session.[5]

For numeric attribute h, the first task is to sequence target tuple $t_1,\cdots,t_N$ based on the size of h. According to definition 2.1, only if $\left|h(t_i)-h(t_j)\right|\le\sigma_h$, the similarity between two target tuples is nonzero value. In this manner, when calculating the similarity between the current scanning tuple and other tuples, because the target tuples are sequenced based on h, other tuples which are actually needed to calculate are the tuples meeting the condition $\left|h(t_i)-h(t_j)\right|\le\sigma_h$, so the total number is limited.

Therefore, $v^f \otimes v^g$ can be divided into two parts to work out, one relies on the current scanning tuple, and the other depends on the related statistics of previously scanned tuples. The similarity between t_i and t_j only needs to calculate once because $Sim(t_i,t_j)=Sim(t_j,t_i)$. During the process of scanning target tuples, for the current scanning tuple t_i, $\eta(i)$ indicates the minimum index number of scanned tuple t_j, which meets $\left|h(t_i)-h(t_j)\right|\le\sigma_h$ before t_i. Thus, $\eta(i)$ increases or unchanged with the increasing of i. During this scanning process, two pointers are maintained, one points to t_i, the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

other points to $t\eta(i)$. All target tuples that meet the condition $|h(t_i)-h(t_j)| \le \sigma_h$ can be got through a single scan. Thus, the following is obtained:

$$v^f \cdot v^h = 2\sum_{i=1}^{N}\sum_{j=\eta(i)}^{i-1} Sim_h(t_i,t_j) \cdot Sim_f(t_i,t_j)$$
$$= 2\sum_{i=1}^{N}\sum_{j=\eta(i)}^{i-1} \{1-[h(t_i)-h(t_j)]\} \cdot [\sum_{k=1}^{l} f(t_i)p_k f(t_j)p_k]$$
$$= 2\sum_{i=1}^{N}\sum_{k=1}^{l} f(t_i)p_k \cdot [1-h(t_i)][\sum_{j=\eta(i)}^{i-1} f(t_j)p_k]$$
$$+ 2\sum_{i=1}^{N}\sum_{k=1}^{l} f(t_i)p_k[\sum_{j=\eta(i)}^{i-1} h(t_j)f(t_j)p_k]$$

In the scanning process, $f(t_i)p_k \cdot [1-h(t_i)]$ and $f(t_i)p_k$ are obtained, at the same time, statistics $\sum_{j=\eta(i)}^{i-1} f(t_j)p_k$ and $\sum_{j=\eta(i)}^{i-1} h(t_j)f(t_j)p_k$ can be worked out through the scanned tuples.[7]

Similarity between two numeric attributes h and g can be worked out through completely different methods. Suppose target tuples are sequenced according to value h. In accordance with definition 2.1, targets tuples meeting the condition of $|h(t_i)-h(t_j)| \le \sigma_h$ and $|g(t_i)-g(t_j)| \le \sigma_g$ are a small part in practical application, then the following method can be adopted to work out $v^f \cdot v^g$.

When dealing with the sequenced target tuples, a search tree is used to conserve all the target tuples which meet the condition of $|h(t_i)-h(t_j)| \le \sigma_h$. This tree is sequenced and indexed based on the value of g. When tuple t* meets the condition and is scanned out, it will be inserted into search tree. All the tuples t meeting the condition of

$$|h(t^*)-h(t)| \le \sigma_h, |g(t)-g(t^*)| \le \sigma_g$$

in the tree can be found out. In this manner, the similarity between two numeric attributes can be worked out by just dealing with this search tree.[8]

The above method improves the computational efficiency of attributes' similarity, however, when the number of target tuples is larger, the method needs further optimized. In fact, $Sim(f,g)$ is two vectors' cosine value based on Euclidean distance, so the optimization can be got by using triangle inequality. The distance between two vectors is

$$|V^f - V^g| = \sqrt{|V^f|^2 + |V^g|^2 - 2|V^f||V^g| Sim(f,g)}$$

For any three vectors f, g and h, V^f, V^g, V^h are vectors based on Euclidean distance, according to triangle inequality

$$|V^f - V^g| \ge \left||V^f - V^h| - |V^h - V^g|\right|$$
$$|V^f - V^g| \le |V^f - V^h| + |V^h - V^g|$$

Before calculating the similarity between f and g, the range of their similarity can be determined by using the similarity of them between other attributes.

Multi-relational attributes and links can be selected by using heuristic search. The multi-relational attributes of user clustering objects are in the first stage. Then, according to the above attribute similarity calculation method, adjacent multi-relational attributes are searched repeatedly. If the similarity of an attribute exceeds a specified threshold, the attribute will be treated as the one to be selected. Assign weights for those selected attributes, and the weight is the degree of relevant to clustering objects. If the weight of the first selected attribute is 1, then the weight of the following searched attributes is the mean of its similarity and initial attribute.

After searching multi-relational attribute, it also needs to determine what relation in the entire database relation diagram to choose, thus, target tuples and their link information will be used. Multi-relational attribute set is determined, so relation set T is the relation of those attributes.

3. MULTI-RELATIONAL CLUSTERING BASED ON LINKS BETWEEN ATTRIBUTES AND TARGET TUPLES

Multi-relational attribute set F and relation set T can be obtained through multi-relational attribute search algorithm. Comparing with objects in different clusters, there is a certain relation within objects in the same cluster, and similarity degree between them is higher. Therefore, it can be concluded that similarity probability of objects without relations must be lower than those with certain relations. Starting from target relation of relation set T, tuple relation diagram G can be obtained. Thus, when calculating similarity between target tuples, the first task is to find whether two target tuples has some relations in diagram G. If there is a related relation, then tuple similarity can be worked out through multi-relational attribute set F, if not, then the similarity between them is regarded as 0.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Suppose after searching $F=\{f_1,\cdots,f_L\}$, each attribute has a weight $f_i.weight$, then, the similarity between any two target tuples t_1 and t_2 is defined as : *if* t_1 *has relations with* t_2

$$Sim(t_1,t_2)=\frac{\sum_{i=1}^{L} Sim_{f_i}(t_1,t_2)\square f_i.weight}{\sum_{i=1}^{L} f_i.weight}$$

Otherwise

$$Sim(t_1,t_2)=0.$$

In this definition, there is a certain connection between t_1 and t_2, which shows in database tuple relation diagram as follows: (1) if there is a route from t_1 to t_2 or t_2 to t_1, then there is a relation between them; (2) if there are same tuples among the tuples linked by t_1 and t_2 along a path, then there is a relation between them; (3) if t_1 and t_2 along a path is linked by the same tuple, then there is a relation between them. Thus, the similarity of target tuples is completed.

CLARANS algorithm is chosen to cluster target tuples. The reason why choose CLARANS is that CLARANS is based on CLARA algorithm and enjoys a preferable scalability on comparatively large data set. CLARANS combines sampling technique and PAM together. At any time, it does not limit any sample but randomly select a sample at every step. The process of clustering can be described as a diagram search. In the diagram, each node is a potential solution, in other words, it is a combination of k representative objects. A clustering result gained by replacing a representative object is regarded as the neighbor of the current clustering result. The number of randomly tried neighbor is limited by a parameter defined by the user. If a better neighbor is found, that is to say it has a better Square-error value, then CLARANS moves to this neighbor node, and the process restarts, otherwise, the current clustering achieve local optimum. If a local optimum is found, CLARANS will start to find a new local optimum from the randomly selected node. CLARANS enjoys a better scalability for multi-relational data set, thus it is the ideal selection algorithm.[10]

4. CLUSTERING ALGORITHM OF TWO-STAGE MULTI-RELATIONAL DATE

In clustering of multi-relational data, object set $X=\{X_1,\ldots,X_m\}$ is given and

$$X_1=\{x_{11},x_{12},\cdots,x_{1n_1}\},\cdots,X_m=\{x_{m1},x_{m2},\cdots,x_{mn_1}\}$$

indicates there are m objects of different categories in X, the number of each category is $n_i(1\le i\le m)$; X_iA is the attribute set of object X_i in each category, the relation among objects in X is $R=\{R_{intra},R_{inter}\}$, among them, $R_{intra}=\{R_1,\cdots,R_m\}$ indicates relations within categories, in other words, the relations among objects of the same categories, while R_{inter} indicates relations between categories, that is the relations among objects of different categories. In all relations, some can be directly obtained by the given data or relation diagram of database, which can be called as explicit relation. Some can be obtained through deep analysis of the data, which can be called as implicit relation. In TSMR algorithm, firstly the given data should be analyzed to get the existed explicit and implicit relations, and those relations are divided into relations within categories and relations between categories. Then, in the first stage, according to the object attribute and information of relations within categories, the objects of each category will be clustered roughly, in this stage, any relation clustering algorithm can be adopted. At the second stage, taking every category cluster of clustering result in the former stage as a new object, according to relations within categories, category clusters of different categories and close relations are combined together to complete hybrid clustering of objects in various categories. [11]

In the first stage, agglomerative hierarchical clustering algorithm is used respectively for each category of object clustering. First, suppose every object is a category cluster, second, calculate similarity of any two ones and combine the closest two category clusters, and repeat this process till no similar one or achieve a specified number. The similarity in clustering is defined as the following: suppose x_{ij},x_{ik} are objects of category X_i, considering the features of objects attribute and relation within categories, the similarity of x_{ij},x_{ik} is defined as:

$$Sim(x_{ij},x_{ik})=\alpha Sim_A(x_{ij},x_{ik})+\beta Sim_{intra}(x_{ij},x_{ik})$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Among this, Sim_A is attribute similarity, Sim_{intra} is similarity of relation within categories, and α, β are weights of each similarity.

The attribute can be classified into numerical attribute and category attribute. Generally, numerical attribute needs standardization. Attribute similarity Sim_A of object x_{ij}, x_{ik} is defined as the sum of difference of attributes, that is

$$Sim_A(x_{ij}, x_{ik}) = \sum_{r=1}^{p} \left| x_{ij}^r - x_{ik}^r \right| + \lambda \sum_{r=p+1}^{N} \delta(x_{ij}^r, x_{ik}^r)$$

Here x_{ij}^r, x_{ik}^r represent r^{th} attribute of objects x_{ij}, x_{ik}, N is the number of attributes, and suppose the attribute from the first to the p^{th} are numerical attribute, and the attributes from $(p+1)$ to N th are category attribute; $\delta(a,b)$ is difference function, when a equals b, the value of $\delta(a,b)$ is 0, otherwise it is 1. Sim_{intra} is the similarity of relation within categories. If the two objects are relations within categories, then the value of Sim_{intra} is 1, otherwise it is 0. When calculating similarity of two category clusters, make sure that attribute similarity is the mean of attribute similarity of objects within the two category clusters, and relation similarity is ratio of the number of relation within categories and the total number of objects in the two category clusters. Suppose c_{ij}, c_{ik} are two category clusters of category X_i, then the similarity between them is

$$Sim(c_{ij}, c_{ik}) = \alpha \cdot \underset{\forall x_{ij} \in c_{ij}, \forall x_{ik} \in c_{ik}}{avg} [Sim_A(x_{ij}, x_{ik})] + \beta \cdot \frac{2\left| R_{intra}(c_{ij}, c_{ik}) \right|}{\left| c_{ij} \right| + \left| c_{ik} \right|}$$

Here C_{ip}, C_{jq} are category clusters of the clustering results in the first stage of category X_i, X_j, $\left|C_{ip}\right|, \left|C_{jq}\right|$ represent the number of objects, $\left|R_{inter}(C_{ip}, C_{jq})\right|$ indicates numbers of relation within categories in C_{ip}, C_{jq}, $\left|X_i\right|, \left|X_j\right|$ are total number of objects in category X_i, X_j, and $\left|R_{inter}(X_i, X_j)\right|$ is the total number of relation within categories of X_i, X_j.

In agglomerative hierarchical clustering algorithm, when the two most relevant category clusters are combined together, the similarity related to new category clusters should be recalculated. At this moment, make sure that similarity is the maximum of similarity between elements among the two category clusters, that is

$$Sim_{inter}(C_i', C_j') = \max_{\forall C_{ip}' \in C_i', \forall C_{jq}' \in C_j'} [Sim(C_{ip}', C_{jq}')]$$

Here C_i', C_j' are two category clusters in the second stage, C_{ip}', C_{jq}' are elements in category clusters, at the initial condition, every element is a category cluster of the clustering result in the first stage.[12]

Two-stage multi-relational data clustering algorithm uses two stages to complete the process of clustering based on in-depth analysis on objects and their relations (including relations within categories, relations between categories, explicit relation and implicit relation). Comparing with a single clustering, it is better to improve algorithm efficiency; in addition, the in-depth analysis promotes cluster quality.

5. APPLICATION OF FUZZYCLUSTERING IN SUBAREA LOAD FORECASTING

It can consider taking subdivision clustering method for n administrative region, which can subdivide the first $i(i = 1, 2, \cdots, n)$ administrative region into m_i pieces[13], so there is small $\sum_{i=1}^{n} m_i$ area totally. It regards environmental factors (output, population, etc.) affecting the load growth as index to $\sum_{i=1}^{n} m_i$ small region fuzzy clustering and clusters into r classes in order to achieve fuzzy clustering of the geographical space. Economic conditions and the law of the load growth are relatively close to each class in the small area. After clustering using same parameters are more reasonable in the same class region.

Let it have n classification object, m factor index, consisting of the following matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

Calculate the average of the first $k(k = 1, 2, \cdots, m)$ factors: $\overline{x_k} = \frac{1}{n}(x_{1k} + x_{2k} + \cdots + x_{nk})$ (1)

Calculate standard deviation:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$S_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2} \quad (2)$$

Calculate standardized value: $x'_{ik} = \frac{x_{ik} - \overline{x_k}}{S_k}$ (3)

Calculate standard value: through the above changes, if Matrix elements value can not be guaranteed in the closed interval $[0,1]$,you can consider using extreme standardization formula to further control the range of matrix elements, calculating the standard value. Concrete forms of extreme standardization formula are as follows:

$$\tilde{x}_{ik} = \frac{x'_{ik} - x_{k\min}}{x_{k\max} - x_{k\min}} \quad (4)$$

$$\tilde{x}_{ik} = \frac{x_{k\max} - x'_{ik}}{x_{k\max} - x_{k\min}} \quad (5)$$

To structure fuzzy relationship matrix, basing on various standardized data of each classified object calculated its similarity degree is need, which was called calibration. Cosine method, similar coefficient method, Euclidean distance method and the method of absolute value subtrahend are commonly used calibration methods. Calibration methods of absolute value subtrahend are as follow:

$$r_{ij} = \begin{cases} 1, i = j \\ 1 - c\sum_{k=1}^{m}|x_{ik} - x_{jk}|, i \neq j \end{cases} \quad (6)$$

Where (6): parameters c should be appropriately selected to ensure $0 \le r_{ij} \le 1$.

When selecting the calibration method, carefully analyze the inherent characteristics of the researched questions and select one or a few more appropriate calibration method. Similar relation of each element is reflected as it and at the same time building the fuzzy similarity matrix r on the matrix x is as follows:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

In general, the fuzzy similar matrix R is just to meet reflexivity and symmetry, but transitivity does not meet, and you can use square method to calculate the transitive closure $t(R)$ .By the transitivity on theorem of fuzzy relations, composite operation process can be applied to the fuzzy similar matrix R:

$$R \to R^2 \to R^4 \to R^8 \to \cdots\cdots$$

and a Positive Integer n exists all the time to let $R^{2^n} = R^{2^{n-1}}$ and get the transitive closure $t(R) = R^{2^{n-1}}$, which is the fuzzy equivalent matrix to be solved.

In accordance with classification standards and classification requirements, use dynamic cluster analysis method in fuzzy mathematics theory, select the appropriate, intercept corresponding cut matrix and get a concrete classification result. By structuring fuzzy statistics you can test whether the number of categories is appropriate and concrete method is as follows:

$$F = \frac{\sum_{j=1}^{r} n_j \left\|\overline{x^{(j)}} - \overline{x}\right\|^2 /(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n} n_j \left\|x^{(j)} - x\right\|^2 /(n-r)} \quad (7)$$

Value F is Large, indicating that the distance between classes is large and the difference between classes is relatively large, at the same time samples are more compact and classification results are better in the same class. Value F obtained provides evidence for determining the appropriate classification by comparing the value F with:

$$F_\alpha(r-1, n-r)(\alpha = 0.05).$$

6. CONCLUSION

In a word, clustering is an important data analysis skill, while traditional data mining is usually assumed that data is composed of the same kind, mutually independent entities. However, in fact, there are a lot of relations among data. The attributes of various types of entity are the same, and the entities are interconnected through various relations. Thus, if the feature of multi-relation of data is ignored, correct conclusion will not be achieved through traditional data mining method. Therefore, there is a need to handle multi-relational feature of data in order to improve the accuracy of data mining result. Currently, multi- relation clustering analysis is becoming one of the important research methods to solve the multi-relation mining problems in mathematics.

ACKNOWLEDGEMENTS

This paper is supported by Scientific and Technological Research Project of Institutions of Higher Education in Hebei Province. Z2012054

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] Imas Sukaesih Sitanggang, Razali Yaakob, Application Of Classification Algorithms In Data Mining For Hotspots Occurrence Prediction In Riau Province Indonesia, Journal of Theoretical and Applied Information Technology, Vol. 43. No. 2,2012, pp. 214 - 221

[2] Aimin Yang, Guanghua Zhao, Yuhuan Cui, Jingguo Qu, "The Improvement of Parallel Predict-Correct Gmres(m) Algorithm and its Application for Thin Plate Structures", *Journal of Computers*, Vol 5, No.10 ,2010, pp.1614-1619.

[3] T. Balakumaran, I.L.A. Vennila, C. Gowri Shankar， Detection of Microcalcification in Mammograms Using Wavelet Transform and Fuzzy Shell Clustering， International Journal of Computer Science and Information Security, IJCSIS, Vol. 7, No. 1, January 2010, pp. 121-125

[4] Aimin Yang, Chunfeng Liu, Jincai Chang and Li Feng, "TOPSIS-Based Numerical Computation Methodology for Intuitionistic Fuzzy Multiple Attribute Decision Making", *Information-an International Interdisciplinary Journal*, Vol. 14,No.10 pp.3169-3174.

[5] Surjeet Kumar Yadav, Saurabh Pal， Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification， World of Computer Science and Information Technology Journal WCSIT, Vol. 2, No. 2, 2012, pp.51-56

[6] Aimin Yang, Chunfeng Liu, Jincai Chang, Xiaoqiang Guo, "Research on Parallel LU Decomposition Method and It's Applica tion in Circle Transportation",*Journal of Software*, Vol 5, No.11,2010, pp.1250-1255.

[7] Brijesh Kumar Bhardwaj, Saurabh Pal，Data Mining: A prediction for performance improvement using classification， (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011, pp.136-140

[8] Lan Wang，GEE analysis of clustered binary data with diverging number of covariates， Annals of Statistics 2011, Vol. 39, No. 1,pp. 389-417

[9] D. Sornette, G. Ouillon， Dragon-kings: mechanisms, statistical methods and empirical evidence， Eur. Phys. J. Special Topics 205, 2012, pp.53-64

[10]Yun Peng, Hongxin Wan, A Fuzzy Evaluation Algorithm Of E-Commerce Customers Based On Attributes Reduction, Journal of Theoretical and Applied Information Technology, Vol. 44. No. 2,2012,pp. 253 - 258

[11] Zhong Zhishui, Wang Gang，The Research On Data Stream Clustering Algorithm Based On Active Grid-Density， Journal of Theoretical and Applied Information Technology， Vol. 44. No. 2 – 2012,pp. 209 - 214

[12] David S. Matteson, Mathew W. McLean, Dawn B. Woodard, Shane G. Henderson,Forecasting emergency medical service call arrival rates,Annals of Applied Statistics, Vol. 5, No. 2B, 2011,pp.1379-1406

[13]Zhou Qian,Han Pu,Zhai Yongjie,Data processing and experimental research on load forecasting,Computer Engineering and Applications,2010,pp.193-195