Classification of High-dimensional Data Clustering Based on Rules Mining Research

Maosen Xia
Dept. of Statistics of Applied Mathematics
Anhui University of Finance and Economics, Bengbu, China
Email:xmsen2000@163.com

Lingling Jiang
Dept. of School of Accountancy
Anhui University of Finance and Economics, Bengbu, China
Email:liliave@163.com

Yumei Wang
Dept. of Statistics of Applied Mathematics
Anhui University of Finance and Economics, Bengbu, China
Email:0517@sina.com

***Abstract*—on the classification of high-dimensional data clustering analysis, traditional similarity index and dimension reduction based on clustering analysis method is hard to avoid "dimension disaster" problem or sampling errors. Therefore, on the basis of choosing the most sub space of the rough set theory, the article directly make a research of the classification of high dimensional data clustering theory mode through to the "equivalence relation" rule mining. Besides, through the China mobile company five cities sampling data of the loss of cell phone users, we has carried on the empirical test and a better clustering results are obtained. In the comparison of K-Means, Two-step and Kohonen methods of clustering, In this paper, classification of high-dimensional data clustering method based on equivalence relation in the type definition, rule mining, the number of iterations which has unique advantages and variable selection.**

***Index Terms*—equivalence relation; rule mining; classification of high-dimensional data; clustering**

I. INTRODUCTION

In the rapid development of information technology, high-speed data volume expansion, increasingly rich data types, and rising data management and analysis demand today, people often face is no longer the traditional sample data, but the vast amounts of high dimensional overall. When the dimension increase, the volume of space improve quickly and thus the available data become very sparse. Sparse for any requirement have statistical significance of the method is a problem, in order to obtain accurate and reliable results in statistics, the amount of data used to support the results needed usually as exponentially with the increase of dimensionality, Which formed the "dimension disaster" (curse of dimensionality) problem. If the observed data as a set of points in high dimensional space, the dimension is higher, the data will be more incline to the border; the distance between any two observations will also tend to be the same threshold. This will lead to the effectiveness of the distance clustering method. According to the problem of "the curse of dimensionality" in theory of classification of data clustering analysis, there are basically two kinds of solution in the existing research results: The one solution is building a new similarity index of thin or phase to reduce the computational complexity. For example, information entropy, similarity and rank effect can be used [1-4]. However, in high-dimensional space , all data are sparse, from many Angles is not similar, so commonly used data organization strategy and cluster becomes very inefficient, " the curse of dimensionality " problem is still can't completely avoid [5]. Second, make the dimension reducing firstly and then clustering. In order to prevent too high dimension lead to hiding clustering, firstly to choose the appropriate subspace, and then make the subspace clustering under the low dimensional [6-7]. But it still uses the distance clustering, you need to sampling to reduce the size of the distance matrix in order to improve the calculation efficiency, obviously sampling design will directly affect the correctness of the results.

To solve the "the curse of dimensionality" problem of high-dimensional categorical data, the classification of high dimensional data clustering theory mode through to the "equivalence relation" rule mining is proposed in this paper. "Equivalence relation of rough set" is essentially the same or similar values on the object in the set of properties to define, with the clustering analysis of the basic ideas happens to have the same view, but don't have to calculate the similarity or thin index [8-10]. This nature makes classification rule mining method based on the equivalence relation can be very good for high-dimensional categorical data analysis. In general, the

doi:10.4304/jsw.9.5.1113-1120

method is based on "equivalence relation "of rough set rules mining, selecting the most sub space, then make the classify of large data clustering. The methods can directly make a clustering analysis for high-dimensional categorical data, avoiding the sample data deviation in sampling, meanwhile the clustering effect is clear and easy to implement. In the example analysis of five cities in the China mobile company mobile phone users ' losing sample data, a better clustering results are obtained.

II. "EQUIVALENT RELATION" CLUSTERING BASED ON ROUGH SETS THEORY AND STATISTICAL UNDERSTANDING

A. Rough set based on Data Description

In the rough sets, can be any classified data set is defined as $D = \{U, A, V, f\}$. Among them, $U = \{x_1, x_2, ..., x_N\}$, as discourse domain, is a finite set of objects; $Q = \{q_1, q_2, ..., q_k\}$ as the description to the objects of the attributes finite set ; V= $U_{q \in Q} V_q$ is constituted by all the attributes of a K-dimensional range space; f: $U \times Q \to V$ is $U \times Q$ to V mapping, while the object is on the property q value function denoted by $f\{x, q\} \in V_q$.

In accordance with the defined problem, or the needs of the research, normally we only need to consider part of the attribute set, or select the subset of attributes. $A \subset Q$ as the scope of the study, the information system is reduced to $D_A = \{U, A, V_A, f\}$.

B. Equivalence Relations

Any subset of domain $X \subset U$ called concepts, and the concept of any arbitrary collection called knowledge. Formed on the rough set theory can only study in U of the division knowledge, that is to the division of knowledge of $p = \{X_1, X_2, ..., X_n\}$, to be satisfied: $X_i \subseteq U$, $X_i \neq f$; $X_i \cap X_j = f$,for $i \neq j$,i,j=1,2,…,n; $U_{i=1}^{n} X_i = U$.

In $D=\{U, Q, V, f\}$, for $\forall A \subseteq Q, A \neq f, \exists x_i, x_j \in U (i \neq j)$ if any

$$f(x_i, q) = f(x_j, q) \ \forall q \subseteq A \tag{1}$$

Says in the D, x_i and x_j （$i \neq j$）in the attribute subset. A are equivalent, denoted as $x_i E x_j$, E_A is called an equivalence relation on the domain U is derived by subset A, abbreviated as E. Easy to know the equivalence relation E meet reflexivity, symmetry, and transitivity. According to the equivalence between E, which concerning the domain can be divided into an equivalence class $U / E = \{X_1, X_2, ..., X_n\}$ Due to the any attribute value (set) (combination) can be divided in theory domain, and get the equivalence class, so that any attributes (set) can form equivalence relation, and in the D, all the set of equivalence relation is called the family of equivalence relation.

If the high-dimensional categorical data as the data points set in high dimensional space, dividing the data form with different attribute sets can be understood as the data point set the projection to the corresponding set of properties of subspace. Thus equivalence relation derived from equivalent classes can be interpreted as in its corresponding set of properties of subspace clustering of data.

C. Core and Reduction

In D={*U, Q, V, f*}, if $A' \subset A \subset Q$ it exists, and $A' \neq f$. Meet

$$E_A = E_{A'} \tag{2}$$

A' as a reduction of *A*. Because of the arbitrariness of loophole set, *A*'s reduction is not unique, the set of all reduction, denoted as *Y*(*A*), and the intersection of all reduction as the nuclear O (A). Obviously there are rules:

$$O(A) = \cap Y(A) \tag{3}$$

Visible, the reduction is a division of the unchanged under the premise of ability, Use the smaller attribute subset to generate equivalence relation; core is the equivalence relations that are generated by all reduction of common attribute subset. So the nucleus is the most basic data of division of knowledge, reduction is in the nuclear division, on the basis of further refinement of knowledge increment.

According to the relationship between core and reduction, not only the reduction can derived kernel, can also the kernel generated by the reduction. If a subset $A' \subset A$ of attributes can be specified for the analysis is particularly important, as A' is the kernel, and find all included reduction which included A' , until meet specific analysis to the date.

This inspiration we can start from a particular attribute, under a certain standard, the equivalence relation gradually generate multiple attributes, and select from it, this process is called equivalence relation rules generation. This process can also be understood as high dimensional data space, namely should check the set of attributes is to keep the most basic data classification information of minimum dimensions set of dimensions, reduction by the corresponding nuclear generating set of attributes is based on the lowest dimension space basis, to further refine the classification of data sets the son the spatial dimension set.

D. Statistical Understanding of Model

From a statistical perspective to understand, high-dimensional data classification is a multivariate general observation data matrix. The essence of the domain of rough set is a sample of the overall sample, is an object, attribute set is the observation variable set, attribute range space is the observation variable set the range space, the mapping function is the value function of the samples in observation. While the equivalence relation is the grouping variable (set), equivalence class is based on

grouping variables of the sample. Because the equivalence relation is meet reflexive, symmetric and transitive, so there is no requirement for grouping variable sequence.

Specifically , the discourse domain which is formed of object collection in the rough set $U=\{x_1,x_2,...,x_N\}$ amount to a sample in Multiple population, the attribute set Q=$\{q_1,q_2,...,q_k\}$ is the set of observed value , map function $f:U\times Q\to V$ is a sample observation function , the equivalence relation which has s attributes $E=\{q_{E_1},q_{E_2},...,q_{E_s}\}$ indicates equivalence relation $U/E=\{X_1,X_2,...,X_n\}$ indicates subsample set which is made of the sub-aggregate of the grouped variable . As for $X_i\in U/E$,i=1,2,…,*n*, $|X_i|$, the cardinal number of X_i , $|U|$, the cardinal number of U , $\sum_{i=1}^{n}|X_i|=|U|$, then $p_i=|X_i|/|U|$ is the probability of the equivalence class X_i , and $p_i\geq 0$, and $\sum_{i=1}^{n}p_i=1$ .So we can get the probability distribution of equivalence relation E

$$E_{|X|}:\begin{Bmatrix}X_1 & X_2 & ... & X_n\\ p_1 & p_2 & ... & p_n\end{Bmatrix} \tag{4}$$

On this basis, we can use information entropy to describe the equivalence relation's reflectlevel to the discourse domain . Promptly

$$H(E)=\sum_{i=1}^{n}p_i\ln(p_i) \tag{5}$$

By formula (5) we can know , when p_i =0 or 1 , $H(E)=0$; when p_i =1/n（i=1,2,…,n, $H(E)$ gets the maximum value $\ln(n)$, namely $H(E)\in[0,\ln(n)]$. Obvious, entropy of information has reflected the amount of information included in equivalence relation E. The information entropy is greater, the greater fluctuant of the equivalence class. In order to avoid the difference of the information entropy absolute number , we use the information entropy relative number indicates the amount of information included in the E, recorded as the information degree α_E, and

$$\alpha_R=\frac{H(E)}{\ln(n)}\times 100\% \tag{6}$$

The greater the degree of information, the more information into the generated equivalence class average carrying, the equivalence relation of domain division more important. So we can make use of the degree of available information to evaluate the effectiveness of the equivalent relations of classification.

III. Classification of High-dimensional Data Clustering Analysis Model

A. Using the Equivalent Relation Classify

We know that the equivalence relation by division of the domain is actually in the high-dimensional space data point set the projection to the lower dimensional subspace, the corresponding set of attributes is the subspace dimension. So we can use equivalence relation of rough set generation process, looking for different subspace projection. To agree with statement of the rough set theory, the following described only by using equivalence relation and the set of properties, deemphasize subspace projection and subspace dimension set.

Although the theory of equivalence relation can be derived domain division, but the optimal classification of equivalence relation is unknown in advanced , so it is necessary to adopt previously mentioned the equivalence relation of the generated method to search out all possible equivalence relation, and choose the most superior price relations as well as the corresponding classification attribute set .

In data set*D*={*U*,*Q*,*V*,*f*} , $U=\{x_1,x_2,...,x_N\}$, if not null subset $A\subseteq Q=\{q_1,q_2,...,q_k\}$, we can get the cutted data set $D_A=\{U,A,V_A,f\}$, we assume A= $\{q_1,q_2,...,q_k\}$. All contains j （$1\leq j\leq k$ ） attribute of the attribute subset A^j （$1\leq j\leq C_k^j$ ） whose assemblage record A^j , intitule the j factors attribute set of A, obviously $|A_i^j|$=j, $|A_i^j|=C_k^j$. Namely $A_1^j=\{q_1,q_2,...,q_j\}$, $A_2^j=\{q_1,q_2,...,q_{j-1},q_{j+1}\}$,…, $A_{C_k^j}^j=\{q_{k-j+1},q_{k-j+2},...,q_k\}$, $A^j=\left\{A_i^j\,|\,i=1,2,...,C_k^j\right\}$, among that C_k^j is combinatorial number. And recorded the equivalence relation family which derived from the j factors attribute set A^j as R^j , recorded j factors equivalence relation , namely

$$R^j=\left\{E_i^j\,|\,i=1,2,...,C_k^j\right\}\qquad j=1,2,...,k \tag{7}$$

Among that the equivalence relation family is derived from the j factors attribute set A_i^j is $E_i^j=E_{A_i^j}^j\left(i=1,2,...,C_k^j\right)$.

Particularly when $j=1$, we can use $A^1=\{\{q_1\},\{q_2\},...,\{q_k\}\}$ to derive one factors equivalence relation family $R^1=\left\{E_i^j\,|\,i=1,2,...,C_k^j\right\}$, among that $E_i^1=\left\{E_{\{q_i\}}^1\right\}(i=1,2,...,k)$; when $j=k$, $R^1=\left\{\left\{E_1^k\right\}\right\}$, we can only get one k factors equivalence relation , namely the equivalence relation which derived

from $A^k = \{\{q_1, q_2, ..., q_k\}\}$.

If it is set that the criterion for classification evaluation is $\theta(\bullet)$, the termination criteria is $\phi(\bullet)$, then the step j is as follows:

- When j = 1, according to the criterion evaluation $\theta(E_*^1)$ in a dollar of equivalence relation R^1 exported from a dollar set of properties A^1, we can select $E_*^1 \in R^1$ for original classification U / E_*^1, then the corresponding a subset of attributes is $A^1 = \{q_*^1\}$;
- When j = 2, A_*^1 products dual set of properties

$$A_*^2 = \{A_i^2 = A_*^1 \cup \{q_i\} \mid A_*^1 \cap \{q_i\} = f; i = 1,2,...,k\} \quad (8)$$

Therefore the family of binary relation R_*^2 exported from A_*^2 contains (*k*-1) equivalent relationships, noting E_i^2 is the equivalent relationship exported from A_i^2.Then,

$$R_*^2 = \{E_i^2 \mid i = 1,2,...,k-1\} \quad (9)$$

According to the criterion evaluation $\theta(E_*^2)$ in R_*^2, we can select $E_*^2 \in R_*^2$ for the second classification U / E_*^2, then the corresponding dual attribute subsets are $A_*^2 = \{q_*^1, q_*^2\}$;

- When j > 2, the j attribute subsets produced by A_*^{j-1} is

$$A_*^j = \{A_i^j = A_*^{j-1} \cup \{q_i\} \mid A_*^{j-1} \cap \{q_i\} = f; i = 1,2,...,k\} \quad (10)$$

Therefore the j relationship family R_*^j exported from A_*^j contains (k-j+1) equivalent relationships

$$R_*^j = \{E_i^j \mid i = 1,2,...,k-j+1\} \quad (11)$$

According to the criterion evaluation $\theta(E_*^j)$ in R_*^j, we can select $E_*^j \in R_*^j$ for the step j classification U / E_*^j, and then the corresponding *j* attribute subset is $A_*^j = \{q_*^1, q_*^2, ..., q_*^j\}$;

- When j = s, the classification stops when the classification of step *s* （ $2 \le s \le k$ ） meets termination criteria $\phi(E_*^s)$.

B. Using Information Entropy to Evaluate the Effect of Classification

Generating equivalence relation just provides the method of classification, and it can't resolve the question of effectiveness of classification. It's necessary to put forward evaluation criteria and abort conditions of classification for clustering by equivalence relation.

If the degree of information of E α_R is greater, E influencing to divide universe is more important than other equivalence relations, therefore, we can use the degree of information as evaluation criteria for choosing equivalence relation. Generally

$$\theta(E_*^j) = \{E_*^j \in R_*^j \mid \alpha_{R_*^j} = \max\{\alpha_i^j \mid i = 1,2,...,C_k^j\}\}$$

j=1,2,…,s (12)

Here

$$\alpha_i^j = \frac{H(E_i^j)}{\ln(n_i^j)} \times 100\% \; i = 1,2,..., C_k^j$$

j=1,2,…,s (13)

The n_i^j is the equivalence class of cardinality exported from the i equivalent relationship E_i^j in j equivalence relation, namely

$$n_i^j = \left|U / E_i^j\right| = \left|X_1^j, X_2^j, ..., X_{n_i^j}^j\right|$$

j=1,2,…,s (14)

We just consider containing equivalence class family of the last classification of attributes in the step of 'classification produced from the equivalence relation'. It means that except the first classification needs counted the information in a dollar of equivalence relation R^1 with k elements, we can produce the j attribute set A_*^j from A_*^{j-1} in the step j($1 < j \le s$) firstly, then export the corresponding relation of equivalence. Therefore the evaluation criterion can be simplified as follows according to $\left|E_*^j\right| = k - j + 1$.

$$\theta(E_*^j) = \{E_*^j \in R_*^j \mid \alpha_{R_*^j} = \max\{\alpha_i^j \mid i = 1,2,...k-j+1\}\} \quad (15)$$

Average information degree may evaluate the amount of information carried by every single equivalence class because the thermal charge about relation of equivalence is determined by different family attribute set. Then the condition to suspend is "the change degree of information after the first j a classification is less than threshold". So when $1 < j \le s$, we can get

$$\phi(E_*^j) = \{E_*^j \in R_*^j \mid \alpha_{R_*^j} / (k - j + 1) > \xi\} \quad (16)$$

When j=s

$$\phi(E_*^S) = \{E_*^S \in R_*^S \mid \alpha_{R_*^S} / (k - j + 1) \le \xi\} \quad (17)$$

IV. THE LIVING EXAMPLE: MOBILE PHONE USERS' LOSS CIRCUMSTANCE OF CLUSTER ANALYSIS

A. Summary of Data, Sample Selection and Data Pre-processing.

According to China's five cities 195608 mobile phone users (a number as a user)data in June 2012, which contains the following properties: phone number, place of residence, age, marital status, income, education level, gender, family number, opening months, wireless services, basic fee, free part, wireless, electronic payment, package type and whether loss, etc. Among them, the wireless service refers to whether to apply for a wireless transfer service, Basic fee refers to the basic fee of last month, free part refers to the last month limiting the cost of the free service items.

For accurate analysis of erosion and retains the characteristics of mobile phone users, first of all, the two data sets is established $D_t = \{U_t, A, V, f\}$ （t=1,2）,Among them, U_1 is the loss of the user object collection, U_2 to keep user object collection, Q feature attribute set for the user, V for the range of values of the attribute set, f is domain mapping of $U_t \times Q$ to V. Second, after remove mobile phone number and whether loss item attribute to get the user characteristics related to the attribute subset. A, including the place of residence, age, marital status, income, education level, gender, family number, opening months, wireless services, the basic fee, free part, wireless, electronic payment and package types. To process the numeric data at the same time, the minimum value (y - y)/maximum - the minimum value of y(y), y is numeric variables, the last set of properties for classification (see table I), after the reduction of two child data set D_1 and D_2.

TABLE I.
ATTRIBUTE DOMAIN DIVISION

Variable	Attribute domains	[0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1]	
v_1	Free part	L	N	M	H	
v_2	The basic cost	L	N	M	H	
v_3	Age	L	N	M	H	
v_4	Opening months	L	N	M	H	
v_5	Wireless charge	L	N	M	H	
v_6	Income	L	N	M	H	
v_7	The number of households	L	N	M	H	
Variable	Attribute domains	**0**	**1**			
v_8	gender	man	Woman			
v_9	Marital status	married	Unmarried			
v_{10}	Wireless services	yes	no			
v_{11}	Electronic payment	yes	no			
Variable	Attribute domains	**1**	**2**	**3**	**4**	**5**
v_{12}	Places to live in	City 1	City 2	City 3	City 4	City 5
v_{13}	The level of education	Junior middle school and the following	Senior high school	University undergraduate course	master's graduate students	PhD student
Variable	Attribute domains	**1**	**2**	**3**	**4**	
v_{14}	Package type	a	b	c	d	

B. The Clustering Process

D_1 loss of users, for example, $A^1 = \{\{v_1\},\{v_2\},...\{v_{14}\}\}$, first from $A = \{v_1, v_2, ..., v_{14}\}$ to generate a set of properties generated by A^1, $R^1 = \{E^1_{\{v_1\}}, E^1_{\{v_2\}}, ..., E^1_{\{v_{14}\}}\}$ again. The resulting information degree α^1_j （j=1,2） a dollar equivalent relation,Shown in table II in R^1 line, obviously $E^1_{\{v_1\}}$ has the largest information degree is 0.8462, so $A^1_* = \{v_1\}$. Generated by A^1_*, $A^2_* = \{\{v_1, v_2\}, \{v_1, v_3\}, ..., \{v_1, v_{14}\}\}$, and calculate R^2_* degree of information, and choose the maximum attribute information of the corresponding binary set $A^2_* = \{v_1, v_2\}$; Repeat until the satisfied end conditions and iteration in the middle of the process are shown in table II.

If the given threshold $\xi = 0.01$ in suspended condition, the clustering step should be stopped in the step j = 7 (table III). So to get the final classification attribute sets $\{v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}\}$.

From the table III can be seen, along with the expansion of attribute sets, average information degree by the start of rapid decrease to gradually decrease steadily. According to the set threshold, clustering step suspended in step 7, is formed by $E^7_{\{v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}\}}$ information equivalence class degrees for α_{Max} =0.5755, the number of equivalence class for E_a =121.

TABLE II.

S_1 THE INFORMATION DEGREE OF EQUIVALENCE RELATION IN THE FAMILY α

R^1		R^2		…	R^7	
v_1	0.5209	v_{14}, v_1	0.5995	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_4$	0.5361
v_2	0.5866	v_{14}, v_2	0.7336*	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_5$	0.5146
v_3	0.6433	v_{14}, v_3	0.4956	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_6$	0.5742
v_4	0.2821	v_{14}, v_4	0.4917	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_7$	0.5379
v_5	0.4056	v_{14}, v_5	0.4355	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_8$	0.5396
v_6	0.2031	v_{14}, v_6	0.4123	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_9$	0.5284
v_7	0.4030	v_{14}, v_7	0.4099	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}$	0.5755*
v_8	0.2157	v_{14}, v_8	0.3281	…	$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{12}$	0.5051
v_9	0.3985	v_{14}, v_9	0.4255	…		
v_{10}	0.5016	v_{14}, v_{10}	0.5198	…		
v_{11}	0.4356	v_{14}, v_{11}	0.5568	…		
v_{12}	0.1618	v_{14}, v_{12}	0.4685	…		
v_{13}	0.4066	v_{14}, v_{13}	0.5344			
v_{14}	0.8462*					

Notes:* said in the maximum number of degree of information in the R1~R7 equivalence relation.

C. Results of Analysis

In the category property set $\{v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}\}$, v_{14} in the first package type, illustrates the package type is most obvious in the loss of user features, then, in turn, is the basic cost, free part, age, electronic payment, education level and wireless services, and the opening of the corresponding month number, the number of wireless expenses, income, family, gender, marital status, where these individual characteristics are not obvious in the loss of users. Specifically:

TABLE III.

MAXIMUM INFORMATION DEGREE AND AVERAGE DEGREE OF INFORMATION IN S_1

Set of properties	The number of equivalence classes	α_{Max}	An average degree of information
v_{14}	5	0.8462	0.2485
v_{14}, v_2	11	0.7336	0.1526
v_{14}, v_2, v_1	23	0.6569	0.0971
v_{14}, v_2, v_1, v_3	41	0.6155	0.0574
$v_{14}, v_2, v_1, v_3, v_{11}$	66	0.5950	0.0394
$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}$	93	0.5875	0.0135
$v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}$	121	0.5755	0.0021
…	…	…	…

- Basic analysis. According to the classification of $\{v_{14}, v_2, v_1, v_3, v_{11}, v_{13}, v_{10}\}$after 121 classification, but only the first 14 classification were more than 2% in the proportion of users, and 1, 2, 3 of the classification is over 10%,Classification of 14 to 86.93% cumulative accounts for users .Will the rest of the small class as the exception class special treatment. The classifications of 15 clustering results are obtained.
- Hierarchical analysis. In order to make the results of the analysis is more practical significance. Further 121 depending on the type of package will be classified into four categories:

  For users of package a, When the basic fee $v_2 \in L \cup N$ and age $v_3 \in L \cup N$ and electronic payment v_{11}=0 (via electronic payment) and Level of education v_{13} = {1, 2}, the users easy to loss. Or basic fee $v_2 \in L \cup N$ and electronic payment v_{11}=1 (not through electronic payment) and free parts $v_1 \in L \cup N$, the users easy to loss;

  For users of package b, when the basic fee $v_2 \in L \cup N$ and age $v_3 \in L \cup N$ and free parts $v_1 \in L \cup N$, the users easy to loss; or basic fee $v_2 \in L \cup N$ and age $v_3 \in L \cup N$ and

Wireless services v_{10} =0(Applied for wireless services), the users easy to loss;

For users of package c, When the basic fee $v_2 \in M \cup H$ and electronic payment v_{11}=1(not through electronic payment) and free parts $v_1 \in L \cup N$ and Level of education v_{13} ={3,4,5} and Wireless services v_{10} =0 (Applied for wireless services),the users easy to loss; Or basic fee $v_2 \in M \cup H$ and Level of education v_{13} ={1,2} and free parts $v_1 \in L \cup N$, the users easy to loss;

For users of package d, When the basic fee $v_2 = H$ and free parts $v_1 = M$,the users easy to loss; Or basic fee $v_2 \in L \cup N$ and Level of education v_{13} ={3,4,5} and electronic payment v_{10} =1(not through electronic payment),the users easy to loss; Or basic fee $v_2 \in L \cup N$ and age $v_3 \in M \cup H$ and electronic payment v_{11} =1(not through electronic payment) ,the users easy to loss.

- The comparison between Equivalent relation clustering and K-Means、Two-step and Kohonen clustering method. Aiming at the loss of user data in the above example, using K-Means, Two-step and Kohonen method which are applied to clustering analysis of the traditional data mining to clustering analysis again. Four clustering methods relevant results are shown in table IV.

From the table4 above can be seen,four clustering methods have some degree of consistency, such as on a variable selection, both of the “Package type” and the “Free part” are significant variables influencing loss of users. "Basic fee" and "age" of these two variables are also significant variable, In addition to the Two - step clustering results； "Electronic payment", "education level" and "wireless services" in the equivalent relation clustering and Two - step clustering are significant variable. While "live" in four kinds of clustering results are shown as insignificant variable.

However the difference of the four clustering methods is very obvious. First of all, on the number of iterations, based on equivalent relation clustering less than K - Means clustering (7 times and 19, respectively); Secondly, on the final number of clustering, four methods have little in common, But in the type definition and rule mining, Base on Package type, only the equivalent relation clustering can divided the user into four types, then we can continue with the Loss of user clustering analysis and rule mining. In contrast, other three kinds of clustering methods, such as K - Means due to the more important variables influencing classification. And on the type definition and rules mining are relatively difficult. Finally, in the screening of

TABLE IV.
THE COMPARISON BETWEEN EQUIVALENT RELATION CLUSTERING AND K-MEANS, TWO-STEP AND KOHONEN CLUSTERING METHOD

	The number of clusters	The number of itera-tions	significant variable	insignificant variable
Equivalent relation clustering	4	7	Package type, basic cost, Free part, age, e-payment, educational level and wireless service	Opening months, Cost of wireless, income, family size, sex, marital status and place of abode
K-Means	9	19	Free part (1.00), basic charge (1.00), Package type (1.00), marital status (1.00), family size (1.00), age(1.00), Opening months (1.00), sex (1.00), income (0.99)	wireless service (0.00), Cost of wireless (0.00), e-payment (0.00), educational level (0.00) and place of abode(0.00)
Two-step	2	-	Free part (1.00), Package type (1.00), family size(0.96), educational level(1.00), wireless service(1.00), Cost of wireless (1.00) and e-payment (1.00)	basic on(0.32), Marital status (0.85), place of abode(0.00), Age (0.84), Opening months (0.14), sex(0.17), income(0.85)
Kohonen	8	-	Free part (1.00), basic cost(1.00), Package type (1.00), marital status(1.00), family size (1.00), age (1.00), Opening months (1.00), sex(1.00), income(0.95)	wireless service(0.00), Cost of wireless (0.00), e-payment (0.00), educational level(0.00)and place of abode(0.00)

Notes: “significant variable” and “insignificant variable” in the table variable name in brackets after the data for the variable importance degree of classification. Because the equivalent relation clustering method is based on the attribute sets the maximum information degree and the average information of convergence condition to select variables, so there is no list of classification of each variable importance degree of the data. The significant variable in K-Means、Two-step and Kohonen clustering refers to the degree of the importance of this variable to classification(inspection probability)is greater than or equal to 0.9,on the other hand, the “insignificant variables” refers to the degree of the importance of this variable to classification is less than 0.9.

Variables, based on equivalent relation clustering can step by step and in turn out seven important variable to the classification of the loss of users (the property set maximum information degree and average information degree are stable). The package type is the most significant variable, While the other three selected by clustering method, the classification of the impact of important variables, Its importance degree are basically is 1.00, it is difficult To distinguish the important differences in the effect of variables for classification.

V. CONCLUSION

In the classification of high-dimensional data clustering analysis of the existing research results mainly can be classified as a new similarity index and Dimension reduction clustering are two thoughts to clustering. With the clustering of dimension reduction method is different, This paper chose the most sub space by using rough set theory, and on this basis, using equivalence relation classification rules in large data directly, clustering analysis, to avoid the sampling bias due to the sample data. Furthermore, compare with other traditional clustering methods of data mining, such as K-Means, Two-step and Kohonen clustering method. This clustering method which is based on equivalent relation also shows its unique advantages.

ACKNOWLEDGMENT

I would like to show my deepest gratitude to Yumei Wang, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance in every stage of the writing of this thesis. In addition, I shall extend thanks to the aid of Social Science fund of Anhui province and Young Scientist Project of Anhui University of Finance and Economics.

REFERENCES

[1] Sida Huang, Qimai Chen. On Clustering Algorithm of High Dimensional Data Based On Similarity Measurement; Computer Applications and Software; Vol. 26 No. 9 pp 102-105, Sep. 2009.
[2] Changsheng Shao, Wei Lou, Limin Yan. Optimization of Algorithm of Similarity Measurement in High Dimensional Data; Computer Technology and Development; Vol. 21, No. 2, pp1-4, Feb. 2011.
[3] Wei Yang, Kuanquan Wang, Wangmeng Zuo. Neighborhood Component Feature Selection for High-Dimensional Data; Journal of Computers;Vol. 7, No. 1,pp 161-167, Jan. 2012.
[4] Md.Anisuzzaman Siddique, Yasuhiko Morimoto. Efficient k-dominant Skyline Computation for High Dimensional Space with Domination Power Index; Journal of Computers;Vol. 7, No. 3,pp 608-615, Mar. 2012.
[5] Sen Wu, Yufei Ye, Xiaoli Yu. Clustering for High Dimensional Data Based on Extended Set Dissimilarity; Application Research of Computers; Vol. 28, No. 9, pp 3253-3255, Sep. 2011.
[6] Zunren Liu, Gengfeng Wu. Quick Reduction Algorithm for High Dimensional Data Sets Based on Neighborhood Rough Set Model; Computer Science; Vol. 39, No. 10, pp 268-317, Oct 2012.
[7] Yong Zhou, XiaoWeLu i, ChunTian Cheng. Parallel Computing Method of Canonical Correlation Analysis for High-Dimensional Data Streams in Irregular Streams; Journal of Software; Vol. 23, No. 5, pp 1053-1072, May 2012.
[8] GuoYin Wang, QingHua Zhang.; Uncertainty of Rough Sets in Different Knowledge Granularities; Chinese Journal of Computers; Vol. 31, No. 9, pp 1588-1598, Sep. 2008.
[9] Hao Chen, JunAn Yang, ZhenQuan Zhuang. The Core of Attributes and Minimal Attributes Reduction in Variable Precision Rough Set; Chinese Journal of Computers; Vol. 35, No. 5, pp 1011-1017, May 2012.
[10] Lijuan Wang, Chen Wu, Xibei Yang, Jingyu Yang. Neighborhood System Based Rough Set and Covering Based Rough Set; Computer Science; Vol. 40, No. 1, pp 221-224, Jan. 2013.

Maosen Xia(1980-), male, Associate Professor, received the Ph.D. degree in management science and engineering in 2012 from Nanjing University of Science and Technology, and is working in Anhui University of Finance and Economics, research directions include high-dimensional data clustering analysis.

Lingling Jiang(1982-), female, lecturer, research directions include the rule mining of high-dimensional data.

Yumei Wang(1965-), female, Professor, research directions include the rough set theory, intelligence computation and optimal control.