Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 329 – 333

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

# Study of qualitative Data Cluster Model based on Granular Computing

Haiyan Li *, Shen Yang, Hong Liu

*School of Science, University of Science and Technology Liaoning, Anshan 114051, China*

**Abstract**

Granular computing theories in the field of computer are introduced into the statistical analysis of qualitative data, based on the traditional qualitative data analysis methods. Multidimensional qualitative data by use of information system are described, and the mathematical model of qualitative data cluster model based on granular computing is given. The feasibility and the superiority are verified by treating massive data. This method may provide a new train of thought for analysis of large and complex qualitative data.



*Keywords:* Granular computing; information system; cluster analysis; qualitative data.

## 1. Introductions

With the rapid development of the information society, the information need to be processed is increasing, and most of the data are not quantitative, which contains a large amount of qualitative data. The traditional data cluster analysis is based on the geometric distance to measure the similarity of the cluster analysis method, and the distance between the qualitative data is unable to accurately measure, in the face of these mass data, the traditional classification methods are limited, there are a great problems in operational efficiency.

* Corresponding author. Tel.: +86-412-5928133; fax: +86-412-5929900.
*E-mail address:* lhyqmj@163.com.


doi:10.1016/j.aasri.2013.10.048

There is high consistency between description for data in information system (Liu, 2001) and description for qualitative data in nature, while the former is more suitable for large data description. First, qualitative data are represented by information system. Then meaningful variable subsets are determined by using of the concept of granularity and solving the granularity matrix progressively (He, 2005, and Yao, 2000). The purpose of classification for massive qualitative data is realized finally.

From the perspective of the information theory, the information content can be used to measure the size of granular, the relationship between granolas represent the degree of dependence similarity or contain (Luo, 2007). Task of cluster analysis is to find the optimal subset and then realize the cluster of observation object. According to this idea, change of importance degree of cluster variable subset can be used to judge the end of cluster process. The mathematical model of qualitative data cluster model is given based on granular computing, the purpose of classification for massive qualitative data is realized, and the feasibility and the superiority are verified by treating massive data (Pu, 2002).

## 2. Description for information system of qualitative data

For $m$ objects, each of which has $n$ components. Let $X=(x_{ij})_{m\times n}$ be an initial data matrix, in which $x_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$ represents numerical value of the $i$ th object in the $j$ th variable. It can be seen that this data matrix has the same form as information system in rough set, represented by the information system $S=(U,A,V,f)$, where $U$ represents non blank finite set of object, it corresponds $m$ objects in qualitative data, called discourse; $A$ represents nonempty finite set of attribute, it corresponds $n$ variables in multidimensional qualitative data; $V=\bigcup_{a\in A}V_a$, $V_a$ represents the range of attribute $A$, corresponds value set of variables in qualitative data; $f: U\times A\to V$ is information function, it gives a information value to every attribute of each object, that is: $\forall a\in A,\ x\in U,\ f(x,a)\in V_a$, corresponds mapping relation of every object and value of variable in qualitative data.

Qualitative data are described in the form of information system, can not only overcome shortcomings of representing massive data by high dimensional contingency table, but also offer basis for the next researches.

## 3. Qualitative data cluster model based on granular computing

### *3.1. Explanations of Symbols*

$S=(U,A,V,f)$: a information system;
$U=\{x_1,x_2,\cdots,x_m\}$: the set of $m$ objects;
$A=\{a_1,a_2,\cdots,a_n\}$: the set of $n$ variables;
$P\subseteq A$: a subset of variable set $A$;
$V_a$: the range of variable set $A$;
$GD(P)$: granularity of variable subset $P$;
$Sig_{A-\{a\}}(a)$: importance degree of variable $a$ to variable set $A$;
$Sig_{A-\{a\}}(a)=GD(A-\{a\})-GD(A)$, $GD(A-\{a\})=\left|U/A-\{a\}\right|^2/\left|A\right|^2$;
$IND(P)=\{(x,y)\in U\times U\mid\forall p\in P, f(x,p)=f(y,p)\}$: indistinguishable relation caused by variable subset $P$.

For $P\subseteq A$, it is easy to verify that indistinguishable relation $GD(P)$ are equivalent relation in set $U$, and $IND(P)=\bigcap_{p\in P}\theta_p$, if $P\subseteq Q\subseteq A$, then $IND(Q)\subseteq IND(P)$.

When $IND(P)$ are the roughest discourse relation (full relation) $\delta$, $IND(P)=U\times U$, $GD(P)$ is the maximum value 1; while $IND(P)$ are the finest discourse relation (equal relation) $\omega$, $\left|IND(P)\right|=\left|U\right|$,

$GD(P)$ is the minimum value $1/|U|$. Therefore, distinguishing power of variables can be described by granularity of variables, the smaller variable's granularity is, the stronger distinguishing power it will have.

According to viewpoint of particle (Huang, 2005), each variable in qualitative data is indistinguishable relation, they all have some distinguishable ability, so each variable has its granularity, the granularity will be smaller when they are combined with each other, and then variable set of smaller granularity and stronger distinguishable ability will come into being.

### *3.2. Calculation and Construction of Model*

Suppose that there are $m$ objects $U=\{x_1,x_2,\cdots,x_m\}$, and $n$ variables $A=\{a_1,a_2,\cdots,a_n\}$. If division of variable set $A$ to $U$ is $U/A=\{X_1,X_2,\cdots,X_s\}$, the problem need to solve is to find a best variable subset $B\subseteq A$, such that the division of subset $B$ to $U$ is $U/B=\{Y_1,Y_2,\cdots,Y_k\}$ ($k<s$).

In the progress of calculating best variable subset, variable subset $B_1=\{a_k,a_l\}$ is obtained at first by calculating granularity matrix of variable $A$ step by step, such that it satisfies:

$$\max_{1\le i,j\le P} GD(a_i,a_j)=GD(a_k,a_l)$$

until the decrement of importance degree of subset after merging sort reach a prescribed critical value $\delta$, that is:

$$Sig_{A-\{B_1\}}(B_1)-Sig_{A-\{B_2\}}(B_2)\le\delta$$

the cluster stopped.

Concrete steps are as follows:

step1: calculate granularity matrix of one dimensional subset of cluster variable set $A$ after screening.

step2: if $\max_{1\le i,j\le P} GD(a_i,a_j)=GD(a_k,a_l)$, then get subset $B_1=\{a_k,a_l\}$, and calculate $Sig_{A-\{B_1\}}(B_1)$.

step3: given critical value $\delta$. If $\Delta=\max\{Sig_{A-\{a_k\}}(a_k),Sig_{A-\{a_l\}}(a_l)\}-Sig_{A-\{B_1\}}(B_1)=0$, then cluster progress stopped in the previous step, and let variable be the highest importance degree as cluster variable; if $\Delta\le\delta$, it indicates that the correlation between $a_k$ and $a_l$ is not marked, stop cluster at this time; if $\Delta>\delta$, then get $B_2$ after calculating granularity between subset $B_1$ and other subsets, and calculate $\Delta=Sig_{A-\{B_1\}}(B_1)-Sig_{A-\{B_2\}}(B_2)$, until it satisfies the requirement of end of cluster.

When the above-mentioned model is carrying out, 2 aspects must be paid more attention:

First, there may be not only one potential important factor behind cluster variable subset which are look for in the process of cluster, that is there may be 2 or more important structures appear in cluster, in that case, cluster process of qualitative data is the same as the description above, while difference is that there may be not only one cluster subset at last.

Second, $\delta$ defines the minimum value of the importance degree of cluster variable subset, the maximum value is 1, considering setting the critical value $\delta$ according to the percentage of importance degree, for instance, importance degree of 5% is 0.05, importance degree of 1% is 0.01, similar to the significance level setting, $\delta$ can be set by different level.

### *3.3. Analysis of Example*

Experiment data used in here come from Bishop, 1976, which were the survey of the preference of detergent products. Set $U=\{x_1,x_2,\cdots,x_{1008}\}$, which takes 1008 informants as observation objects; variable set $A=\{a_1,a_2,a_3,a_4\}$, where $a_1$ represents softness, range $V_{a_1}=\{1(\text{soft}),2(\text{midle}),3(\text{hard})\}$; variable $a_2$ represents if it had used trademark $M$, range $V_{a_2}=\{1(\text{yes}),2(\text{no})\}$; variable $a_3$ represents the temperature of

water, range $V_{a_3} = \{1(\text{high}), 2(\text{low})\}$ ; variable $a_4$ represents trademark liked by informants, range $V_{a_4} = \{1(\text{trademark } X), 2(\text{trademark } M)\}$ . Equivalence relation formed by variable set classifies informants into 24 categories; see concrete data in table 1.

Table 1. Survey data of the preference of detergent products

| $U$ | $A$ | | | |
|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
| $x_1$ | 1 | 2 | 2 | 2 |
| $x_2$ | 3 | 1 | 1 | 2 |
| $x_3$ | 2 | 1 | 1 | 1 |
| ... | ... | ... | ... | ... |
| $x_{1008}$ | 1 | 2 | 1 | 1 |

Calculate one-dimensional granularity matrix of variable set $A = \{a_1, a_2, a_3, a_4\}$ , see computing result in table 2.

Table 2. One-dimensional granularity matrix of variable set $A$

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $a_1$ | 1 | 0.001 | 0.004 | 0.000 |
| $a_2$ | 0.001 | 1 | 0.001 | 0.016 |
| $a_3$ | 0.004 | 0.001 | 1 | 0.003 |
| $a_4$ | 0.000 | 0.016 | 0.003 | 1 |

From table 2 can be know that $\max\{GD(a_i, a_j)\} = GD(a_2, a_4) = 0.016$ , thus subset $B_1 = \{a_2, a_4\}$ , and calculate its importance degree, then

$$Sig_{A-\{B_1\}}(B_1) = 0.894$$

$$\max\{Sig_{A-\{a_2\}}(a_2), Sig_{A-\{a_4\}}(a_4)\} = 0.899 ,$$

So

$$\Delta = \max\{Sig_{A-\{a_2\}}(a_2), Sig_{A-\{a_4\}}(a_4)\} - Sig_{A-\{B_1\}}(B_1) = 0.005$$

Given $\delta = 0.02$ , the cluster progress stopped because $\Delta < \delta$ .

Cluster variable subset $B_1 = \{a_2, a_4\}$ , variable $a_2$ represents detergent product if it had used trademark $M$ , variable $a_4$ represents liked trademark. As can be seen that variable subset represents spending habits of the informants. Therefore informants can be classified according to spending habits: 1) informants who had used detergent products of trademark $M$ , liked trademark is $X$ ; 2) informants who had not used detergent products of trademark $M$ , liked trademark is $X$ ; 3) informants which had used detergent products of trademark $M$ , liked trademark is $M$ ; and 4) informants who had not used detergent products of trademark $M$ , liked trademark is $M$ . Facts also indicate consumers pay much more attention to brands than softness and temperature of water, such classification fits the reality.

In conclusion, equivalence relation formed by variable set is classified into data of 1008 informants of 24 categories, determined variable subset $B_1 = \{a_2, a_4\}$ which is meaningful to cluster by calculating granularity matrix, then calculate importance degree after merging sort and do comparative analysis, define threshold value $\delta$ , classify informants into 4 categories from spending habits, thus stop the whole cluster progress. Further show the feasibility of qualitative data cluster model based on granular computing, it can not only manipulate massive data but also increase the computational efficiency.

## 4. Conclusion

Qualitative data are described by using information system, a mathematical model of qualitative data cluster is given based on granular computing, and a new method of determine best cluster variable subset upon the basis of analyzing variable subset is proposed. This cluster method not only fits cluster which has one important classification structure data, but also fits cluster which has multiple important classification structure data. Examples show that this model is highly effective in the face of massive data, and provides a new idea for analysis of huge and complex qualitative data.

## Acknowledgements

The work was supported by Natural Science Foundation of Liaoning Province under Grand 20102097.

## References

[1]Liu Qing. Rough sets and rough reasoning. Beijing: Science Press, 2001: 23-40.
[2]He Ming, Feng Boqin, Ma Zhaofeng, Fu Xianghua. Rough set clustering algorithm based on entropy and information granularity. Journal of Xi′an Jiaotong University, 2005, 39(4): 343-346.
[3]Yao YY. Granular computing: basic issue and possible solutions. Proceedings of the 5th Joint Conference on Information Sciences, 2000: 186-189.
[4]Luo Min. Granular computing and its current status of research. Computer and Modernization, 2007, 1: 1-5.
[5]Pu Dongbo, Bai Shuo, Li Guojie. Principle of granularity in clustering and classification. Chinese Journal of Computers, 2002, 25(8): 810-816.
[6]Huang Zhaohua, Deng Yixiong. An approach for granular computing and its application. Journal of East China Jiaotong University, 2005, 22(5): 124-127.
[7]YM Bishop, SE Feinberg, PW Holland, Discrete multivariate analysis: theory and practice, MIT Press, Cambridge, Mass., 1976.