A Global Discretization and Attribute Reduction Algorithm based on K-means Clustering and Rough Sets Theory

HU Min
School of Management and Economics, Beijing Institute of Technology, Beijing, 100081, China.
e-mail: humin0516@gmail.com

***Abstract*—The knowledge reduction function of rough sets theory is specific on discrete data, while most attributes of decision tables are continuous. Therefore a global discretization and attribute reduction algorithm is proposed based on clustering and rough sets theory. After comparing different discretization methods, the *k*-means clustering algorithm is used. In order to avoid the shortcomings of *k*-means clustering algorithm, the F—analysis of variance statistics and support strength of condition attributes are introduced to control the discretization effectiveness. A rational clustering number is derived according to the dependency index to meet the prerequisite of the rough set theory. After that, the attributes are reduced by using rough set theory, and decision rules are induced. Lastly an example is proposed to illustrate the feasibility and effectiveness of the algorithm.**

Keywords-Rough set; Decision table; K-means clustering; Discretization; Attribute reduction;

1. INTRODUCTION

Rough sets theory is a useful technique to deal with discrete (qualitative) data, but actually most attributes of decision table are continuous(quantitative). The shortcomings of dealing with continuous attributes limit its application range. Therefore, the continuous data should be substituted for limited semantic variables(or symbols) through the discretization process[1-3]. Then the attributes can meet the need of rough sets theory for attributes reduction and rule induction. The results of discretization depend on different discretization methods, such as demesne knowledge discretization, equidistant division and equal frequency discretization and so on[4, 5]. In consideration of the classification information of attributes, clustering method is used to discrete continuous attributes.

This paper is organized as follows: Firstly, the F—analysis of variance statistics and support strength of condition attribute are introduced to control the discretization effectiveness. Secondly a supervisory mechanism is set up specific to the shortcomings of *k*-means clustering algorithm. Thirdly, a global discretization and attribute reduction algorithm is proposed based on clustering and rough sets theory. The comparison result of the case illustrates that the algorithm is more effective and can be a useful decision support tools.

2. THE ROUGH SETS CONCEPT

Rough Sets theory, first introduced by Pawlak in the year 1982, is a mathematical theory widely used in domains of data mining, decision analysis and pattern recognition and so on[6-8]. Rough Sets theory is good at dealing with problems of uncertainty and identifying cause-effect relationship in databases as a form of database learning[9].

Concept1:Decision table

According to rough sets theory, an information system is represented as follows: $S=\{U,A,V,f\}$, where U and A, are finite, nonempty sets called the universe, and the set of attributes, respectively. $A=C\bigcup D$, in which C represents a finite set of condition attributes and D represents a finite set of decision attributes; $V=\bigcup_{r\in A}V_r$ represents the domain of attributes value, in which V_r is the domain of the attribute $r\in A$; $f:U\times A\to V$ is the total decision function called the information function. The decision table is the tabular form of S.

Concept2: Set approximation

The idea of rough sets actually consists of the approximation of a set by a pair of sets called the lower approximation and the upper approximation.

If $S=\{U,A,V,f\}$ is a decision table, $[x]_R$ represents a set composed of the elements under the equivalent relation of R. Every subset X of the universe U is assigned two sets $R_*(X)$ and $R^*(X)$ called the *R*-lower and the *R*-upper approximation of X, respectively. The *R*-lower approximation objects is defined as $R_*(X)=\{x\in U,[x]_R\subseteq X\}$, where objects unambiguously belong to set X, The *R*-upper approximation objects is defined as $R^*(X)=\{x\in U,[x]_R\cap X\neq\varnothing\}$, where objects possibly belong to set X. The set $BN_R(X)=R^*(X)-R_*(X)$ will be referred to as the *R*-boundary region of X. The R positive region and the R negative region of X is $\mathrm{POS}_R(X)=R_*(X)$ and $\mathrm{NEG}_R(X)=U-R^*(X)$ respectively. If the boundary region of X is the empty set, i.e., $BN_R(X)=\varnothing$, then the set X will be called crisp (exact) with respect to R; in the opposite case, i.e., $BN_R(X)\neq\varnothing$, the set X will be referred to as rough (inexact) with respect to R.

Concept3: Attribute reduction

If $S=\{U,A,V,f\}$ is a decision table, $A=C\bigcup D$, the dependent relationship of decision attributes D to condition attributes C is described by quality of dependency. We say that D depends in degree $\gamma_C(D)$, $0\le\gamma_C(D)\le1$, on C, denoted, if

$$\gamma_C(D)=\mathrm{card}(\mathrm{POS}_C(D))/\mathrm{card}(U), \quad (1)$$

where $\mathrm{POS}_C(D)=\{x\in U,[x]_C\subseteq D\}$.

If $\gamma_C(D)=\gamma_{C-c}(D)$, $c\in C$, then the condition attribute

978-0-7695-3888-4/09 $26.00 © 2009 IEEE
DOI 10.1109/KAM.2009.16

c is unnecessary to the decision attributes set D, otherwise is necessary. If all condition attributes are necessary to the set decision attributes, the set of condition attributes is independent to the set of decision attributes. The significance of c $(c \in C)$ to D is defined as $Sig(c) = \gamma_C(D) - \gamma_{C-c}(D)$. This number, expressing the importance of an attribute in a decision system, is evaluated by measuring the effect of removing the attribute from the table. The process of removing the unnecessary attributes is called attribute reduction.

3. Global Discretization and Attribute Reduction Algorithm

3.1 Discretization based on k-means clustering method

Depending on different measurements of similarity levels, a set of multi-dimensional data samples (n) can be divided into k categories ($k < n$) during clustering process[10, 11]. The ultimate objective is to make samples of same categories have highest similarity, while samples from different categories have the lowest similarity. Clustering algorithms[12] are well developed and can be broadly divided into hierarchy method[13], partition method, fuzzy set theory, density-based and grid-based approaches, etc. K-means clustering is a kind of partition methods[14]. The computational cost of k-means clustering is comparatively shorter and has a linear correction with the size of data sets. One prerequisite of k-means clustering is that the clustering number should be pre-designated. This may affect the effectiveness of attribute reduction subsequently. For these reasons, two auxiliary statistics and variables are introduced to supervise the clustering results.

Normally the ideal effect is that the clustering number is as little as possible on premise of the clustering effectiveness. The clustering results will be quite different when different clustering methods are used. Therefore it is necessary to introduce some variables and standards to verity the effects. There are two rules of analyzing the clustering effects: One for general clustering, the clustering result will be better with a greater distance between categories, and with the higher similarity among the members. The other for clustering according to decision rules and knowledge, the ideal effect is to ensure that the results of clustering can preserve the necessary rules of the raw data and avoid the loss of information.

In order to certify the clustering effect, following parameters are chosen to determine the clustering number.

1)F-statistic of analysis of variance

F-statistic of analysis of variance is defined as

$$F = \frac{MS}{MS'} = \frac{SS/v}{SS'/v'} = \frac{\sum_{i=1}^{k} n_i \left| \bar{X}_i - \bar{X} \right|^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_k} \left| x_{ij} - x_i \right|^2 / (n-k)}, \quad (2)$$

where MS and MS' represent the standard deviation from inter-class and intra-class respectively, SS and SS' represent the sum of squares of mean deviation from inter-class and intra-class respectively, v and v' denote degree of freedom from inter-class and intra-class respectively; k is the number of class , n is the number of samples, $\bar{X}$ denotes the centre of all elements, n_i is the number of the class i, $\bar{X}_i$ denotes the centre of class i. The relationship of distance between inter-class and intra-class is illustrated from the formula. Larger F-statistic indicates more rational clustering results.

2) Support strength of condition attributes

If $a \in C$ is a condition attribute, $y \in D$ is a decision attribute, $S_a(y)$ represents the support subset of y, in which $S_a(y) = \cup_{W \in U/y} \left(\cup_{V \in U/a, V \subseteq W} V \right)$, and $Spt_a(y) = \mathrm{card}(S_a(y)) / \mathrm{card}(U)$ is the support strength of a to y. Larger support strength indicates better category capacity.

3.2 Global discretization and attribute reduction algorithm

On the premise of keeping the relationship between the condition attributes and the decision attributes, the range of value region of continuous attributes is divided into a number of small areas by setting up several demarcation points. Thus the key problem of discretization of continuous attributes is to determine the number and the position of demarcation points.

Therefore, combining the merits of k-means clustering under supervision with the merits of discrete number processing of rough sets theory, a global discretization and attribute reduction algorithm based on clustering and rough sets theory is proposed. The algorithm can also derive the decision rules from the continuous valued decision table.

The algorithm is represented blow:

Step 1: Suppose the number of all attributes in condition set is m, we assign the number of clustering category k_d to the condition attributes d $(d = 1, 2, \cdots m)$, and set the acceptance value β of the dependency of the decision set to the condition set;

Step 2: Set the initialization of $d = 1$ to begin the clustering process from the first condition attribute;

Step 3: Divide the condition attribute d to k_d categories using clustering method;

Step 4: Let $d = d + 1$, goes back to the step 3 until $d = m$ to cluster all the condition attributes to k_d categories;

Step 5: Caculate the dependency value of the decision set to the condition set $\gamma_C(D)$, if $\gamma_C(D) \geq \beta$, the class number of the condition attributes can be decreased by 1 gradually; while if $\gamma_C(D) < \beta$, the dependency could not meet the requirement. So that the class number of all condition attributes should be increased by 1, let $k_d = k_d + 1$, $d = 1, 2, \cdots m$, go back to step 2;

Step 6: Set the initialization of $d = 1$ to begin the decreasing of class number from the first condition attribute;

Step 7: Let $k_d = k_d - 1$ and calculate the dependency value $\gamma_C(D)$ again. If the dependency doesn't change, let $k_d - 1$ be the class number; while if it changes smaller,

the class number should keep still;

Step 8: Let $d = d + 1$, go back to step 7 until $d = m$;

Step 9: Go back to step 6 until no class number can be reduced. If the class number of an attribute is 1, the attribute is an unnecessary attribute.

When the amount of data is comparatively less, we can assign $\beta = 1$ hoping all the decision rules reserved. Relaxation of restrictions is considered only if there are large amounts of data.

4. COMPARISONS OF RESULTS FOR DIFFERENT CASES

4.1 Case Discription

Decision-making table is structured as follows: The universe $U = \{E_1, E_2, \cdots, E_m\}$ represents the set of candidate virtual enterprises supporting certain kind of resource product. The condition set is $\boldsymbol{C} = \{C_1, C_2, C_3, C_4\}$, which represents four indicators of the quality respectively: C_1 represents the quality certification; C_2 represents the technology; C_3 represents the percent of pass; C_4 represents the repair rate. Decision-making set is $\boldsymbol{D} = \{B\}$, which represents the final quality grade of each enterprise on basis of expert advice and practical experience.

The quality decision table of certain kind of virtual enterprises is as follows:

TABLE I. QUALITY DECISION TABLE OF A KIND OF VIRTUAL ENTERPRISE

U	C				D
	C_1 (%)	C_2	C_3 (%)	C_4 (%)	B
E_1	45.00	8.00	89.00	15.00	2
E_2	95.00	9.00	87.00	20.00	2
E_3	70.00	7.50	79.00	23.00	1
E_4	90.00	9.00	92.00	5.00	3
E_5	86.00	7.00	81.00	27.00	1
E_6	50.00	6.50	83.00	30.00	1
E_7	75.00	7.00	82.00	22.00	2
E_8	55.00	7.50	85.00	26.00	1
E_9	80.00	7.00	80.00	28.00	1
E_{10}	79.00	7.00	82.00	24.00	1
E_{11}	48.00	8.00	93.00	12.00	3
E_{12}	82.00	8.50	89.00	8.00	3

4.2 K-means clustering and attribute reduction with pre—designated k

4.2.1 K-means clustering and validation with pre—designated k

Let $k_d = 3$, the discretization results can meet the prerequisite of rough set theory using k-means clustering. The clustered quality decision table of virtual enterprises is as follows:

TABLE II. CLUSTERED QUALITY DECISION TABLE OF VIRTUAL ENTERPRISES

U	C_1	C_2	C_3	C_4	B	U	C_1	C_2	C_3	C_4	B
E_1	1	1	3	1	2	E_7	2	3	2	3	2
E_2	3	2	2	1	2	E_8	1	1	2	3	1
E_3	2	1	1	3	1	E_9	2	3	1	3	1
E_4	3	2	3	2	3	E_{10}	2	3	2	3	1
E_5	3	3	1	3	1	E_{11}	1	1	3	1	3
E_6	1	3	2	3	1	E_{12}	2	2	3	2	3

The F-statistics value of C_3 is $28.719 >> 1$, the corresponding probability is $p \ll 0.001$. Therefore the clustering process is effective.

4.2.2 Attribute reduction and rule induction

Through the induction of condition set, the unnecessary attribute with significance value of 0 is excluded. The set C' composed of attributes left is called the reduction of condition set C. The significance of condition attribute C_1、C_2、C_3、C_4 is:

$Sig(C_1) = 1/12$, $Sig(C_2) = 0$, $Sig(C_3) = 1/12$, $Sig(C_4) = 0$.

Therefore the technology and the repair rate are the unnecessary attributes, while the quality certification and the percent of pass are cores of decision attributes. The reduced condition set is $C' = \{C_1, C_3\}$. By simplifying the reduced condition set and combining the same rules, the candidates choosing rules is obtained as follows:

TABLE III. CANDIDATES CHOOSING RULES (1)

rules	Condition attributes	Decision attribute	rules	Condition attributes	Decision attribute
1	C_1 (2), C_3 (1)	B=1	6	C_1 (3), C_3 (2)	B=2
2	C_1 (3), C_3 (1)	B=1	7	C_1 (1), C_3 (3)	B=2
3	C_1 (1), C_3 (2)	B=1	8	C_1 (1), C_3 (3)	B=3
4	C_1 (2), C_3 (2)	B=1	9	C_1 (3), C_3 (3)	B=3
5	C_1 (2), C_3 (2)	B= 2	10	C_1 (2), C_3 (3)	B=3

The fourth and the fifth rules, and the seventh and eighth rules of the table are obscure. That means some information has lost during the clustering process. That's also why $\gamma_C(D)$ is less than 1. Therefore we need go on using the global discretization and attribute reduction to compute this case.

4.3 Global discretization and attribute reduction process

When the class number is $k_d = 3$, the dependency $\gamma_C(D) = 8/12 < \beta = 1$ can't meet the terminal condition. Thus we should increase the class number of all condition attributes. The new decision table with four class attributes is obtained when we let $k_d = 4$. This new decision table can match the prerequisite with $\gamma_C(D) = 1 = \beta$, and the class number of each attribute can be decreased.

(1) First, we reduce the class number of C_1. When $k(C_1) = 3$,

$k(C_2) = 4$, $k(C_3) = 4$, $k(C_4) = 4$,

$\gamma_C(D) = \mathrm{card}(\mathrm{POS}_C(D)) / \mathrm{card}(U) = 5/6 < \beta$.

This indicates that we can't reduce C_1 and should keep $k(C_1) = 4$ still;

In a similar way when $k(C_2) = 3$, $k(C_1) = 4$, $k(C_2) = 4$, $k(C_4) = 4$, the dependency $\gamma_C(D) = 1$ can meet the prerequisite. Thus we can reduce the class number of C_2 to $k(C_2) = 3$;

In the same way, we keep the number of C_3 still and reduce the class number of C_4 to $k(C_4) = 3$.

(2) Keep on reducing the class number until

$k(C_2)=1$ ， $k(C_4)=1$。

The final decision table is as follows:

TABLE IV. DECISION TABLE BASED ON GLOBAL DISCRETIZATION AND ATTRIBUTE REDUCTION ALGORITHM

U	C_1	C_2	C_3	C_4	*B*	*U*	C_1	C_2	C_3	C_4	*B*
E_1	1	1	1	1	2	E_7	3	1	2	1	2
E_2	2	1	1	1	2	E_8	1	1	2	1	1
E_3	3	1	3	1	1	E_9	4	1	3	1	1
E_4	2	1	4	1	3	E_{10}	4	1	2	1	1
E_5	4	1	3	1	1	E_{11}	1	1	4	1	3
E_6	1	1	2	1	1	E_{12}	4	1	1	1	3

The final results are:

$\gamma_C(D)=1$ ， $\gamma_{C-\{C_1\}}(D)=\gamma_{C_3}(D)=5/12$ ，

$\gamma_{C-\{C_2\}}(D)=1$ ， $\gamma_{C-\{C_3\}}(D)=\gamma_{C_1}(D)=0$ ，

$\gamma_{C-\{C_4\}}(D)=1$， The final significance of each condition attributes is: $Sig(C_1)=7/12$ ， $Sig(C_2)=0$ ，

$Sig(C_3)=1$， $Sig(C_4)=0$. The reduced condition set is also $C'=\{C_1,C_3\}$.

On the whole, the global discretization and attribute reduction algorithm is better than the former ones through the comparison of different indexes. By simplifying the condition attributes and combining the similar rules, the candidates choosing rules containing 10 clear and concise causalities is obtained. The final rules set can avoid the confusion of the former results:

TABLE V. CANDIDATES CHOOSING RULES (2)

rules	Condition attributes	Decision attribute	rules	Condition attributes	Decision attribute
1	C_1 (3)， C_3 (3)	*B*=1	6	C_1 (2)， C_3 (1)	*B*= 2
2	C_1 (4)， C_3 (3)	*B*=1	7	C_1 (3)， C_3 (2)	*B*=2
3	C_1 (4)， C_3 (2)	*B*=1	8	C_1 (2)， C_3 (4)	*B*=3
4	C_7 *(1)*， C_9 *(2)*	*B=1*	9	C_1 (1)， C_3 (4)	*B*=3
5	C_1 (1)， C_3 (1)	*B*=2	10	C_1 (4)， C_3 (1)	*B*=3

5. CONCLUSIONS

A global discretization and attribute reduction algorithm based on *k*-means clustering and rough sets theory has been proposed and experimented in the numerical example. The F—analysis of variance statistics and support strength of condition attribute have been introduced to compare and verify the discretization effectiveness. And the dependency defined in rough sets theory has been introduced to determine a rational clustering number. Through attribute reduction, useful knowledge and rules can be induced and refined to the support decision maker.

The experiment of the numerical example has confirmed that the method is effective in decision-making data analysis and extracting central and concise rules. Rough sets theory can simplify the indicator system through its attribute reduction function. It can not only reduce the complexity of follow-up calculation, but simplify the process of data collection and save large numbers of human and material resources. This should be built on the condition of covering space cases to meet the prerequisite of rough set theory.

The accuracy and adaptability of the refined rules will be verified in further research. The expansion and optimization of rules repository will be studied further so as to use this method to a wide range of areas.

ACKNOWLEDGMENT

The first author thanks the National Natural Science Foundation, research grant no. 70841008.

REFERENCES:

[1] Deng Wu, Yang Xin-hua, Zhao Hui-min. An algorithm for decision table reduction based on rough set and hybrid clustering method[J]. Journal of Dalian Jiaotong University,2008,29(3):86-90. (in chinese)

[2] Guan Yan-yong, Wang Hong-kai, Wang Yun, et al. Attribute reduction and optimal decision rules acquisition for continuous valued information systems [J]. Information Sciences, 2009, 179(17):2974-2984.

[3] Guan Yan-yong, Wang Hong-kai, Wang Yun, et al. Attribute reduction and optimal decision rules acquisition for continuous valued information systems [J]. Information Sciences, 2009, 179(17):2974-2984.

[4] YAN Zhi-jun, ZHANG Yue-jun. Decision table reduction based on cluster analysis[J]. Transactionsof Beijing Instituteof Technology , 2006,26(3):256-259. (in chinese)

[5] LIU Ye-zheng, JIAO Ning, JIANG Yuan-ehun. Study on comparison of discretization algorithms of continuous attributes[J]. Application Research of Computers, 2007, 24(9): 28-33. (in chinese)

[6] [Pawlak Z. Rough sets [J]. Int J of Computer and Information Science, 1982, 11(5): 341-356.

[7] Pawlak Z. Rough sets-theoretical aspects of reasoning about data [M]. Kluwer: Kluwer Academic pub., 1991.

[8] LIU Qing. Rough Set and Rough Reasoning[M].Beijing: Science Press, 2001. (in chinese)

[9] Shen Li-xiang, Francis E. H. Tay, Qu Liang-sheng, et al. Fault diagnosis using rough sets theory[J]. Computers in Industry, 2000,43(1):61-72.

[10] Amitava Roy, Sankar K. Pal Fuzzy discretization of feature space for a rough set classifier[J]. Pattern Recognition Letters, 2003, 24(6):895-902.

[11] XIANG Xin-jian, Stolle. M. An algorithm of discretization of continuous attributes in rough sets based on cluster[J]. Journal of Zhejiang University of Science and Technology,2003,15(3):154-157.

[12] E Xu, GAO Xuedong, CHEN Yi, et al. Clustering Algorithm Based on Rough Set[J]. Computer Engineering, 2007, 33(4):14-16.(in chinese)

[13] Zhang Qing-hui, Mu Xiao-dong, Li Yi, et al. A global discretization algorithm based on hierarchical method[J]. Microcomputer Information, 2009,25(5-3):213-215. (in chinese)

[14] Wen Feng, Chen Zong-hai, Zhuo Rui, et al. Reinforcement learning method of continuous state adaptively discretized based on k-means clustering[J]. Control and Decision,2006,21(2):143-146.(in chinese)