The Difference Degree of Condition Attributes and Its Application in the Reduction of Attributes

Taorong Qiu*
Department of Computer, Nanchang University, Nanchang 330031, China
Email: taorongqiu@163.com

Yuyuan Lin and Xiaoming Bai
Department of Computer, Nanchang University, Nanchang 330031, China

***Abstract*—The attribute reduction based on rough set usually reflects the essence of information table and it is the key content of rough set theory. With the analysis of the difference degree between equivalence classes or tolerance classes based on rough set theory, a measure for the importance of attributes was introduced. And based on the measure, a method of the reduction of attributes was proposed and a corresponding reduction algorithm was designed. This method is applicable to a complete information system and incomplete information system. In the algorithm, the measure was used as heuristic factor to find a reduction subset. And also, the approximation quality of classification was applied to evaluate the quality of the reduction subset. Finally, test and comparison results show that the proposed method is feasible.**

***Index Terms*—Rough Set, Attribute Reduction, Importance of Attributes, Difference Degree of Sets**

I. INTRODUCTION

The attribute reduction based on rough set usually reflects the essence of an information table and it is the key content of rough set theory[1-2]. Nowadays, the main research directions of the reduction of attributes focus on the more effective attribute reduction algorithms, the optimal feature subset, reducing the time complexity of algorithms and so on[3-6]. Many researchers have proposed a lot of reduction algorithms from different point of views. Miao proposed a heuristic algorithm for reduction of knowledge[7] and a knowledge reduction algorithm under graph view[8]. Liang[9] presented an information quantity based heuristic algorithm for reduction of attribute. Wang[10-11] presented an algorithm of reduction based on the information entropy. Generally, it is shown that the existing attribute reduction algorithms have some disadvantages such as the time performance, the quality of a reduction subset and so on. For generating a reduction of attributes rapidly and accurately on data sets with core attributes, this paper introduces a difference degree between equivalence classes or tolerance classes based on rough set theory, proposes a measure for the importance of attributes. And based on the measure, a method of reduction of attributes is proposed and a corresponding reduction algorithm is designed. This method is applicable to a complete information system and incomplete information system. In the algorithm, the measure was used as heuristic factor to find a reduction subset. And the approximation quality of classification was applied to evaluate the quality of the reduction subset.

Manuscript received February 2, 2011; revised April 1, 2011; accepted May 4, 2011.

* Corresponding author. Tel.: +86 0791 8305570.

II. DIFFERENCE DEGREE OF CONDITION ATTRIBUTES AND EVALUATION OF REDUCTION SUBSETS

A. Difference Degree and Importance of Condition Attributes

(1) Similarity and difference degree between sets

Firstly, the similarity and difference degree between sets are introduced as follows.

Definition 1 Given two sets X , Y , the similarity degree between X and Y denoted by $Sim(X,Y)$ is defined as follows. $Sim(X,Y)=|X\bigcap Y|/|X|$. And the difference degree denoted by $Dif(X,Y)$ between X and Y is defined as follows. $Dif(X,Y)=(|X|-|X\bigcap Y|)/|X|$, where $|X|$ denotes the cardinality of the X .

Obviously, $Sim(X,Y)=1-Dif(X,Y)$.

(2) The similarity and difference degree between equivalence classes in a complete information system

Secondly, the similarity and difference degree between equivalence classes in a complete information system are described

Let $S=(U,A,V,f)$ be a complete information system with decision attributes., namely a complete decision system, where $A=C\bigcup D$ is a nonempty attribute set, C denoting condition attributes and D standing for decision attributes, and U is a nonempty finite set of objects, such that $f:U\to V_a$ for any $a\in A$, where V_a is called the value set of a . Let $IND(P), P\subseteq C$ be a binary indiscernibility relation,

doi:10.4304/jcp.7.5.1087-1091

namely equivalence relation. The relation $IND(P)$ constitutes a partition of U, which we denote by $U/IND(P)$, for short U/P.

For any $IND(P)$, a equivalence class of $U_i \in U$ is denoted by $S_P(U_i)$ .Similarly, for any $IND(D)$, a equivalence class of U_i is denoted by $S_D(U_i)$.

Definition 2 A similarity degree between equivalence classes is defined as follows. $Sim(S_P(U_i), S_D(U_i)) = |S_P(U_i) \bigcap S_D(U_i | / |S_P(U_i)|$. And a difference degree between equivalence classes is defined as follows:

$$Dif(S_P(U_i), S_D(U_i)) = (|S_P(U_i)| - |S_P(U_i) \bigcap S_D(U_i|)/|S_P(U_i)|$$

Based on the difference degree, a measure for the importance of attributes is defined as follows.

Definition 3 The importance degree of any condition attribute subset $P \subseteq A$ with respect to decision attributes D is defined as follows.

$$SIG_{IND}(P,D) = (1/|U|)\sum_{i=1}^{|U|}(Dif(S_{C-P}(U_i), S_D(U_i)))$$

(3) The similarity and difference degree between tolerance classes in an incomplete information system

Thirdly, a similarity and difference degree between tolerance classes in an incomplete information system are described

Let $S = (U, A, V, f)$ be an incomplete information system with decision attributes, namely an incomplete decision system.

Let $SIM(P), P \subseteq C$ be a binary tolerance relation. The relation $SIM(P)$ constitutes a covering of U, which we denote by $U/SIM(P)$.

For any $SIM(P)$, a tolerance class of $U_i \in U$ is denoted by $SIM_P(U_i)$. Similarly, for any $SIM(D)$, a tolerance class of U_i is denoted by $SIM_D(U_i)$.

Definition 4 A similarity and difference degree between tolerance classes is defined as follows. $U/SIM(P) = \{SIM_P(U_1), SIM_P(U_2), ..., SIM_P(U_{|U|})\}$. And a difference degree between tolerance classes is defined as follows.

$$Dif(SIM_P(U_i), SIM_D(U_i)) = (|SIM_P(U_i)| - |SIM_P(U_i) \bigcap SIM_D(U_i|)/|SIM_P(U_i)|$$

(4) A measure for the importance of attributes

Definition 5 an importance degree of any condition attribute subset $P \subseteq A$ with respect to decision attributes D is defined as follows.

$$SIG_{SIM}(P,D) = (1/|U|)\sum_{i=1}^{|U|}(Dif(SIM_{C-P}(U_i), SIM_D(U_i)))$$

Obviously, the less powerful classing ability of an attribute subset $P \subseteq C$, the smaller its importance degree. On the contrary, the more powerful classing ability of an attribute subset, the bigger its importance degree.

B. Evaluation of reduction subsets

It is one of importance steps for evaluating a reduction subset in the process of generating attribute reduction subsets. Generally, two indexes are used to evaluate whether the reduction is good or bad. One is the degree of importance of an attribute reduction subset for classification, generally the bigger the better. The other is the degree of attribute dependency between reduction subsets, which can describe superfluous attributes, generally the smaller the better.

In this paper, the approximate quality of classification is used as a measure to evaluate whether an attribute reduction subset is good or bad.

Definition 6 Let $S = (U, A, V, f)$ be a complete information system with decision attributes. The approximate quality of classification on $P \subseteq C$ is defined as follows.

$$\gamma_P = \sum_{i=1}^{|U/D|} \left|\underline{S_P}(X_i)\right| \Big/ |U|$$

Where $X_i \in U/D$, and $\underline{S_P}(X_i)$ is the lower approximation of X_i.

Definition 7 Let $S = (U, A, V, f)$ be an incomplete information system with decision attributes. The approximate quality of classification on $P \subseteq C$ is defined as follows.

$$\gamma_P = \sum_{i=1}^{|U/SIM(D)|} \left|\underline{SIM_P}(X_i)\right| \Big/ |U|$$

where $X_i \in U/SIM(D)$, and $\underline{SIM_P}(X_i)$ is the lower approximation of X_i.

Obviously, $\gamma_P \in (0,1)$

III. AN ALGORITHM FOR ATTRIBUTE REDUCTIONS BASED ON THE DIFFERENCE DEGREE

In an incomplete decision system, because there exist some null values, many attributes may become core attributes. So the positive region will change irregularly. In our method, an attribute reduction begins with an empty set without computing the core attributes and ends in satisfying with the given constrain conditions by adding the attribute with the most importance degree in current computing step. The algorithm for generating a

reduction of attributes based on the measure for the importance of attributes is described as follows.

Algorithm 1: An algorithm for an attribute reduction based on the difference degree

Input: A complete decision system or an incomplete decision system $S=(U,A,V,f)$, where $A=C\bigcup D$, and the threshold value ε.

Output: reduction subset $\mathrm{Re}d$

Algorithm description:

Step 1: Initialization: Let $\mathrm{Re}d=NULL$, $P=C$;

Step 2: Judging whether the given decision system is complete or not;

Step 3: Computing the difference degree of attribute subset P;

For a complete decision system, computing the difference degree according to the formula $Dif(S_P(U_i),S_D(U_i))$;

For an incomplete decision system, computing the difference degree according to the formula $Dif(SIM_P(U_i),SIM_D(U_i))$

Step 4: Computing the importance degree of each attribute in P;

For a complete decision system, computing the importance degree according to the formula $SIG(S_P,S_D)$

For an incomplete decision system, computing the importance degree according to the formula $SIG(SIM_P,SIM_D)$

Step 5: For $\forall C_i\in P$, selecting the attribute denoted by C_k with the most importance degree and adding it into $\mathrm{Re}d$, namely $\mathrm{Re}d=\mathrm{Re}d\cup\{C_k\}$, and let $P=P-\{C_k\}$;

Step 6: Computing $\gamma_{\mathrm{Re}d}$. If $\gamma_{\mathrm{Re}d}<\varepsilon$,then continuing step 4, else goto step 7;

Step 7: Outputting $\mathrm{Re}d$;

Step 8: Algorithm ending.

IV. Experiments and Result Analysis

A . Experiment Introduction

(1) Experimental environment

CPU: Intel(R) Core(TM)2 Duo CPU 1.80GHz

Memory:1.00GB

Hard disk:160G

Operation system: Windows XP Professional

Developing environment: using C++ language, running in VC6.0

(2) Contents for testing and comparing

The contents for testing and comparing of the algorithm include two aspects. One is to test the running time of the algorithm. The other is to evaluate the reduction obtained whether it is good or bad.

B. Testing and Comparing of Algorithms in a Complete Decision System

(1)Datasets

Four datasets from UCI have been selected to validate the effectiveness of the proposed algorithm. These data sets only need a small amount of discrete or inherently discrete. Table 1 shows the description of data sets.

(2)Testing results

For comparison of our algorithm (denoted by ARDD)), we select existing four other algorithms for generating reduction sunsets. They are the reduction algorithm (denoted by A1) based on the positive region, the reduction algorithm (denoted by A2) based on the importance degree of attributes, the reduction algorithm (denoted by A3) based on the discernibility matrix and the reduction algorithm (denoted by A4) based on information entropy.

The algorithm A1,A2,A3,A4 and our method respectively are run on each dataset and two results are obtained. One result is running time of each algorithm. The other result is the percentage of the cardinal number of the reduction subset generated to the total number of attributes in the datasets.

The experimental results are the average of algorithm iteration for ten times. Test results are shown in table 2.

TABLE 1. DATASETS WITHOUT NULL VALUES FOR TESTING AND COMPARING OF ALGORITHMS

No.	Name	Sample number	Condition attribute number	Decision attribute number
1	Zoo	72	16	1
2	Vote	83	16	1
3	Soybean	114	35	1
4	Splice	83	60	1

TABLE 2. TESTING RESULTS ON DATASETS WITHOUT NULL VALUES

Dataset No.	Algorithm A1		Algorithm A2	
	Executing Time(ms)	Percentage	Executing Time(ms)	Percentage
1	--	--	13111	56.3%
2	1692	25%	7653	43.8%
3	--	--	18346	31.4%
4	--	--	37717	10%
Dataset No.	**Algorithm A3**		**Algorithm A4**	
	Executing Time(ms)	**Percentage**	**Executing Time(ms)**	**Percentage**
1	795	18.8%	15868	100%
2	1484	37.5%	16186	100%
3	12115	8.6%	18411	80%
4	32415	50%	Overflow error	
Dataset No.	**ARDD**			
	Executing Time(ms)		**Percentage**	
1	829		12.5%	
2	1062		25%	
3	6750		11.4%	
4	16875		5%	

Note :"—" denotes that no result is obtained

From table 2, it is shown that the algorithm A1 don't generate any reduction subset on three datasets. By

analyzing it is known that these datasets don't have any core attributes. So, it is not suitable for the reduction algorithm based on the positive region to generate reduction on a dataset without core attributes.

The algorithm A2,A3 and A4 can obtain reduction subsets on any dataset, but it can not be guaranteed that the cardinal number of the reduction subsets is the smallest in all reduction subsets. Furthermore, the run-time of one of algorithm A2, A3 and A4 becomes more greater than that of the proposed algorithm when the scale of a dataset being bigger.

From the above table, it is shown that the proposed algorithm can not only obtain a reduction subset with the smallest length on four datasets, but have good performance in running time as well when datasets being bigger.

C. Testing and Comparing of Algorithms in an Incomplete Decision System

(1)Datasets

Three data sets from UCI have been selected to validate the effectiveness of the proposed algorithm. Table 3 shows the description of data sets.

TABLE 3. DATASETS WITH NULL VALUES FOR TESTING AND COMPARING OF ALGORITHMS

No.	Name	Sample number	Condition attribute number	Decision attribute number	Loss rate of data
1	Adult	90	14	1	32.5%
2	Horse	120	27	1	19.8%
3	Anneal	195	38	1	61.8%

(2)Testing results

For comparison of our algorithm (denoted by ARDD)), we select existing two other algorithms for generating reduction sunsets. They are the reduction algorithm (denoted by B1) based on the positive region and the reduction algorithm (denoted by B2) based on the discernibility matrix.

The algorithm B1,B2 and our method respectively are run on each dataset and two results are obtained. One result is running time of each algorithm. The other result is the percentage of the cardinal number of reduction subset generated to the total number of attributes in the datasets.

The experimental results are the average of algorithm iteration for ten times. Test results are shown in table 4:

TABLE 4. TESTING RESULTS ON DATASETS WITH NULL VALUES

Dataset No.	Algorithm B1		Algorithm B2	
	Executing Time(ms)	Percentage	Executing Time(ms)	Percentage
1	--	--	2478	35.6%
2	20476	37%	46701	26.3%
3	2369	5.4%	3203	13.2%
Dataset No.	**ARDD**			
	Executing Time(ms)		**Percentage**	
1	1031		21.4%	
2	21812		37.0%	
3	4967		3.3%	

From table 4, it is shown that our algorithm can obtain a satisfied reduction subset and a good running time performance on datasets with null values comparing to other two algorithms. Similarly, it is not suitable for the reduction algorithm based on the positive region to generate reduction on a dataset without core attributes.

V. CONCLUSION AND FUTURE WORK

The attribute reduction based on rough set usually reflects the essence of an information table and it is the core content of rough set theory. In this paper, with the analysis of advantage and deficiency of the existing attribute reduction algorithms, an attribute reduction method based on the difference degree of sets has been proposed In the algorithm, the similar relation and equivalence relation have been combined and the difference degree has been used as heuristic factor to find the reduction set in a dataset, which may be an complete decision system or an incomplete decision system. And also, the approximation classified quality has been applied to evaluate the quality of reduction subsets. Test and experiment results show that the proposed method is feasible. Especially the method can find out reduction subsets rapidly and accurately on data set with a core.

Further work involves such aspects as optimizing the algorithm, testing further in some large datasets, making comparison with other methods in all aspects, and so on.

ACKNOWLEDGMENT

Supported by National Natural Science Foundation of China (#61070139) and the Science and Technology Planning Project of the Education Department of Jiangxi Province in China (GJJ11286,GJJ11037, S00945)

REFERENCES

[1] Q. Liu, *Rough Sets and Rough Reasoning(in Chinese)*. Beijing: Science Press, 2005, pp.11-75.

[2] D. Y. Deng, Research on Data Reduction Based on Rough Sets and Extension of Rough Set Models(in Chinese), Ph.D Thesis, Beijing Jiaotong University, 2007.

[3] D. Q. Miao, G. Y. Wang, and Q. Liu(Eds.), *Granular Computing: Past, Present, and the Future Perspectives(in Chinese)*. Beijing: Academic Press, 2007

[4] G. Y. Wang, *Rough Theory and Knowledge Discovery(in Chinese)*. Sian: Sian Jiaotong University Press, 2001, 133-139

[5] W. X. Zhang, Y. Liang, and W. Z. Wu, *Information System and Knowledge Discovery(in Chinese)*. Beijing: Science Press, 2003,89-96

[6] K. Y. Hu, Y. C. Lu and C. Y. Shi, "Advances in rough set theory and its applications(in Chinese)", *Journal of Tsinghua University (Sci. & Tech.)*, Vol. 41,pp. 125-137, 2001.

[7] D. Q. Miao, G. R. Hu, "A heuristic algorithm for reduction of knowledge(in Chinese)", *Journal of Computer Research &Development*, Vol. 36, pp. 681-684, 1999.

[8] D. Q. Miao, Y. M. Chen, "Knowledge reduction algorithm under graph view(in Chinese)", *ACTA ELECTRONICA SINICA*, Vol. 38, pp.1952-1957, 2010.
[9] G. Y. Liang, K. S. Qu and Z. B. Xu, "Reduction of attribute in information systems(in Chinese)", *Systems Engineering — Theory & Practice*, Vol. 21,pp.76-80, 2001.
[10] G. Y. Wang, H. Yu and D. C. Yang, "Decision table reduction based on conditional information entropy(in Chinese)", *Chinese Journal of Computers*, Vol. 25, pp.759-766, 2002.
[11] G. Y. Wang, "Calculation methods for core attributes of decision table(in Chinese)", *Chinese Journal of Computers*, Vol. 26, pp.611-615 , 2003.
[12] G. J. Huang, S. L. Wang and X. G. Zhang, "Query expansion based on associated semantic space", *Journal of Computers*, Vol. 6, pp. 172-177, 2011

Taorong Qiu received the Ph.D degree in computer application technology from Beijing Jiaotong University, Beijing, China, in 2009. He has been a member of China Rough Sets and Soft Computing Committee. He has (co)authored more than 30 scientific research papers in indexed journals, books, Rough Sets and Soft Computing committee. His research interest covers rough sets, granular computing and knowledge discovery.

Yuyuan Lin received the Diploma in computer engineering degree from the Department of Computer Engineering at the Sanming College, China in 2009. Currently, he is a graduate student with the Department of Computer, at Nanchang University. His research interest includes rough set and modeling for prediction.

Xiaoming Bai received the Diploma in Engineering degree from the Department of Electrical Engineering at the Nanchang University, China in 1986. Currently, he serves as an associate professor with the Department of Computer, Nanchang University. His research interest covers computer networks, data mining.