Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 3, March 2018, pp. 573-578
DOI: 10.23940/ijpe.18.03.p18.573578

A Calculation Method for Dependency Degree of Condition Attribute Set using Discernibility Matrix

Hongchan Li[a], Junxing Liu[b], and Haodong Zhu[a,*]

[a]*School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, 450002, China*
[b]*No.1 Middle School of ZhengZhou, Zhengzhou, Henan, 450007, China*

Abstract

In the process of attribute reduction, the importance degree of a condition attribute is generally measured by means of the dependence degree between the condition attribute and the decision attribute set. If the dependence degree of the condition attribute is 0, we generally think that the condition attribute does not affect the decision results of the decision table and can be directly deleted form the condition attribute set. However, to some extent, it cuts off the connection of the condition attribute and other attributes, resulting in a great loss of valuable information in the decision table. Therefore, based on the fact that the dependency degree of the condition attribute set is more credible than the dependency degree of a single condition attribute, this paper researches the dependency degree of the condition attribute set and puts forward a calculation method for dependency degree of condition attribute set using a discernibility matrix. This paper also presents and proves a theorem to improve the proposed method. The proposed method can quickly get the discernibility matrix and can directly calculate the dependency degree of the condition attribute set. The theoretical analysis and the simulation experiment comparison results all show that the proposed method has better effectiveness and lower time complexity.

Keywords: rough set; attribute reduction; dependency degree; discernibility matrix

(Submitted on July 17, 2017; First revised on October 27, 2017; Second revised on November 21, 2017; Accepted on December 21, 2017)

1. Introduction

Rough set was proposed by Polish scholar Z. Pawlak in 1982 [2] and has been widely used in many fields, such as machine learning [19], data mining [7], fault diagnosis [12], fuzzy control [18], and so on. Its core research content is to reduce knowledge to obtain the most representative condition attribute subset and the simplest set of classification rules according to the condition attribute set, the decision attribute set and the decision object set of decision table or information table [17]. Knowledge reduction includes attribute reduction [14] and classification rule reduction [5]. At present, domestic and foreign scholars mainly study in terms of attribute reduction [6].

In attribute reduction, the importance degree of a condition attribute is generally measured by means of the dependence degree between the condition attribute and the decision attribute set [16]. If the dependence degree of a condition attribute is 1, then the decision attribute set is totally dependent on the condition attribute; if the dependence degree of a condition attribute is 0, then the decision attribute set is totally independent on the condition attribute; if the dependence degree of a condition attribute is in the interval (0,1), then the decision attribute set is partially dependent on the condition attribute. At present, there are many research achievements on dependence degree of a single condition attribute [1,4,10,11]. It is generally acknowledged that the condition attribute with dependence degree 0 does not affect the decision results of the decision table and can be deleted directly from the condition attribute set. However, to some extent, it cut off the connection of the condition attribute and other condition attributes, so that it results in a greater loss of valuable information in the decision table.

* Corresponding author.
E-mail address: zhuhaodong80@163.com

In fact, the dependency degree of the condition attribute set is more credible than the dependency degree of a single condition attribute. It not only reflects the importance degree of condition attributes on the decision results, but also maintains the organic relation between condition attributes. At present, there are many research achievements on dependence degree of condition attribute set [9,13,15,20]. However, these methods are generally generated on the basis of equivalence class, and their time complexity are very high, which is not conducive to solving relevant problems.

Because the discernibility matrix has a great advantage in attribute reduction, it has been widely used in attribute reduction [8]. In this case, this paper studies the dependency degree of condition attribute set based on discernibility matrix, a calculation method for dependency degree of condition attribute set is proposed, and its lower time complexity is proven by mathematical theoretical analysis and its effectiveness is verified with practical examples.

2. Basic theory

Given a decision information system $S=(U,A,V,f)$, U is a non-empty finite set of data objects, $A=C\bigcup D$ is an attribute set, C and D are the condition attribute set and the decision attribute set respectively, $f: U\times A\to V$ is an information function. The basic knowledges related to this paper are as follows [3]:

Definition 1 (Equivalence class). An equivalence class of indiscernibility relation based on the condition attribute set $B\subseteq C$ is denoted by $U\mid B=(X_1,...,X_p)$, X_j is an object set of x_i (x_i is an object in U) that all condition attribute values in B are equal. The equivalence class with indiscernibility relation based on the decision attribute set D is denoted by $U\mid D=(Y_1,...,Y_q)$, among them, Y_j is an object set of x_i (x_i is an object in U) that all decision attribute values in D are equal.

Definition 2 (Positive region). Given $X_i(i=1,2,...,p)$ is an equivalence class based on the condition attributes set $B\subseteq C$, $Y_j(j=1,2,...,q)$ is an equivalence class based on the decision attribute set D. The lower approximation of Y_j is defined by

$$P\_(Y_j)=\cup\{X_i\in U\mid X_i\subseteq Y_j\},\ i=1,2,...,p;\ j=1,2,...,q. \tag{1}$$

Definition 3 B Positive region of D is denoted by

$$POS_B(D)=\cup P\_(Y_j),\ \ j=1,2,...,q. \tag{2}$$

The knowledge of a decision information system can be described by the decision attribute set D, and can also be described by the condition attribute set C. As to which to use, it depends on the dependency relationship between functions in a database.

Definition 4 (Dependency degree). Given $S=(U,A,V,f)$ is a decision information system, $A=C\bigcup D$ is an attribute set. C and D are the condition attribute set and the decision attribute set respectively. The dependency degree that describes the decision attribute set D based on the condition attribute set $B\subseteq C$ is defined by

$$k=\frac{card(POS_B(D))}{card(U)} \tag{3}$$

It is called that D depends on B with the dependency degree $k(0\le k\le 1)$, and is denoted by $B\Rightarrow D$. Among them, $card()$ represents the cardinality of a set.

Definition 5 (Discernibility Matrix). Given $S=(U,A,V,f)$ is a decision information system, $U=\{x_1,x_2,...,x_n\}$ is an object set with n objects, $A=C\bigcup D$ is an attribute set, $C=\{a_1,a_2,...,a_m\}$ is a condition attribute set with m condition attributes, $D=\{d_1,\ldots,d_p\}$ is a decision attribute set with p (p>=1)decision attributes. For any two objects x_i and x_j, the element of discernibility matrix is defined by

$$M_{ij} = \begin{cases} \{a_k \mid a_k(x_i) = a_k(x_j), a_k \in C\}, if \cdots \exists d \in D, d(x_i) \neq d(x_j) \\ \phi, if \cdots \forall d \in D, d(x_i) = d(x_j) \end{cases} \quad (4)$$

Theorem 1. x_i and x_j are objects in decision information system. If any condition attribute set B is included in an element M_{ij} of the discernibility matrix, that is $B \subseteq M_{ij}$, then x_i and $x_j \notin POS_B(D)$.

Proof. According to the definition of discernibility matrix, the decisions attribute values of the object x_i and x_j are different, that is $\exists d \in D, d(x_i) \neq d(x_j)$. But for any condition attribute $a_k \in B$, the attribute values of the object x_i and x_j are equal, that is $a_k(x_i) = a_k(x_j)$. Meanwhile, according to the definition of positive region, x_i and $x_j \notin POS_B(D)$.

Definition 6 (Dependency degree of condition attribute set). Given $S = (U, A = C \cup D, V, f)$ is a decision information system, $B \subset C$ is a condition attribute set which consists of more than one attribute. The dependency degree that describes the decision attribute set D based on the condition attribute set B is called the dependency degree of condition attribute set B.

3. Proposed calculation method for dependency degree of condition attribute set

If more than one dependency degrees of condition attribute are equal in a decision information system, it is difficult to determine which attribute is more important. In this case, in order to distinguish the importance degree of these attributes, we need combine the evaluated attribute with other condition attributes (two or more) and calculate the dependency degree of these condition attributes. At present, there are many research achievements on dependence degree of condition attribute set [9,13,15,20]. These research achievements are generally generated on the basis of equivalence class. The algorithm in literature [15] is a classical and more representative dependency degree evaluation algorithm of this kind, which many other existing degree evaluation algorithms for condition attribute set are based on. We take the algorithm in literature [15] as an example to analyze the performance of this kind of algorithm. The detailed algorithm in literature [15] is described as follows:

```
Begin:// Generate the equivalence class
    for each combination of condition attributes
      for i=1 to n
        for j=i+1 to n
          Compare the values of objects x_i and x_j on the selected combination of condition attributes, and put x_i
          and x_j to an equivalent class if their attribute values are equal;
        end for
      end for
    end for
   nCount=0;
   for each equivalent class
      if the decision attribute values are all equal then
        nCount=nCount+ the number of objects in the equivalent class;
        The dependency degree of condition attribute set=nCount/ Card(U);
    end for
  return The dependency degree of each combination of condition attributes;
End.
```

The time complexity analysis: This algorithm needs construct single attribute strings in the form of multiple condition attributes. If there are m condition attributes and n objects in a decision information system, then it needs $\sum_{k=1}^{m} C_m^k$ add operations. Meanwhile, the equivalent classes of the formed single attribute string also need be generated. The overall time complexity of the algorithm is about $O(n^2 \times m^2)$, which is very high and is not conducive to solving relevant problems.

Because the discernibility matrix has a great advantage in attribute reduction, it has been widely used in attribute reduction [8]. In order to overcome the deficiencies of the above algorithms, this paper proposes a new calculation method for dependency degree of condition attribute set based on discernibility matrix and is described as follows:

Begin:
Generate the discernibility matrix $\boldsymbol{M}$ according to the definition 5;
for any attribute set B
$A=U$;
for each object x_i of A
if there exists an element M_{ij} in the i-th row of the discernibility matrix M and $x_i \subseteq M_{ij}$, then delete x_i and x_j form A;
end for
the dependency degree of condition attribute set B=Card(A)/Card(U);
end for
return the dependency degree of each condition attribute subset.
End.

The time complexity analysis: If there are m condition attributes and n objects in in a decision information system, the time complexity of the proposed method consists of two parts: ① the time complexity $O(m \times n^2)$ for obtaining the discernibility Matrix, and ② the time complexity $O(m^2 \times n)$ for the calculate the dependency degree of each condition attribute subset. So, the total time complexity of this proposed method is $O(m \times n^2 + m^2 \times n)$, which has an advantage over the algorithm of literature [15].

4. Experimental verification

On the experiment platform with 2.40GHz CPU frequency, 8GB RAM and Windows 8.1 operation system, we employ the following six data sets [15] to compare the proposed method with the algorithms in literatures [13,15,20] to validate the performance of the proposed method.

① A decision information table is shown in Table 1, which includes 4 condition attributes, 1 decision attribute and 8 objects.

Table 1. A decision table

U	a	b	c	d	D
x1	2	0	1	0	0
x2	2	1	1	0	0
x3	0	1	1	0	0
x4	1	2	2	0	1
x5	1	2	0	1	1
x6	0	2	2	1	1
x7	0	2	1	1	2
x8	2	2	1	1	2

② Bupa data set in UCI knowledge base provided by University of California which includes 6 condition attributes, 5 decision attributes and 428 data records.

③ Diabetes data set in UCI knowledge base provided by University of California which includes 8 condition attributes, 2 decision attributes and 798 data records.

④ Breast data set in UCI knowledge base provided by University of California which includes 9 condition attributes, 2 decision attributes and 817 data records.

⑤ Segmentation data set in UCI knowledge base provided by University of California which includes 19 condition attributes, 7 decision attributes and 2932 data records.

⑥ Chess End-Game data set in UCI knowledge base provided by University of California which includes 36 condition attributes, 1 decision attribute and 3196 data records.

For the convenient expression, the proposed method in this paper is referred to as "NM", the algorithm in literature [15] is referred to as "LM", the algorithm in literature [13] is referred to as "BM" and the algorithm in literature [20] is referred to as "TM".

We adopt MATLAB 7.0 to implement numerical calculation, select the consumed time (units: s) as the measure of time complexity, and employ the close degree between the calculation result of dependency degree of each selected algorithm and the calculation result of dependency degree of the classical rough set theory [3] as the measure of dependency degree of condition attribute set. In the experiment, each algorithm is performed 10 times, and we take the average time of the 10 records as the experimental results of the consumed time and the average of close degree of the 10 records as the experimental results of the close degree. The experimental comparison results are shown in Table 2.

As seen from Table 2, the close degrees of the four algorithms are approximately equal and are up to 99%. This indicates that the four algorithms can all obtain a better condition attribute subset when they are used in attribute reduction. The consumed time of the four algorithms from large to small is about the proposed LM ≈TM ≈ BM > NM, and the time of NM is far lower than that of the other three algorithms. With the increase of data set scale, the bigger the data set scale is, the more obvious the time advantage of NM is.

Table 2. The experimental comparison results of the four algorithms

Experimental Data Sets	①				②			
Compared Algorithms	NM	LM	BM	TM	NM	LM	BM	TM
Close Degree	100%	100%	100%	100%	99.58%	99.58%	99.47%	99.58%
Consumed Time(s)	1.0877	2.1031	1.8962	1.7015	5.5117	7.9379	7.1495	7.0071
Experimental Data Sets	③				④			
Compared Algorithms	NM	LM	BM	TM	NM	LM	BM	TM
Close Degree	99.37%	99.42%	99.21%	99.32%	99.58%	99.58%	99.58%	99.58%
Consumed Time(s)	8.099	13.7562	12.1029	12.1991	8.5047	14.1135	13.0305	13.1039
Experimental Data Sets	⑤				⑥			
Compared Algorithms	NM	LM	BM	TM	NM	LM	BM	TM
Close Degree	99.72%	99.48%	99.61%	99.59%	99.13%	98.99%	99.10%	99.13%
Consumed Time(s)	15.9188	33.7499	31.9918	32.1752	17.0414	36.0718	35.4729	35.1752

Through careful analysis, we find that the four algorithms are all based on the classical rough set theory and have not deviated from the actual decision results. Therefore, their close degrees are very high and approximately equal. However, because the time complexity of NM is $O(m \times n^2 + m^2 \times n)$, the time complexity of the other three algorithms are all $O(n^2 \times m^2)$. So, in the case of a small amount of data, the consumed time of the four kinds of algorithms has no obvious difference, but with an increase of data set scale, the advantage of the time performance of the proposed method is very obvious.

5. Conclusions

In a decision information table, the condition attribute with dependence degree 0 may have an impact on other condition attributes. In that case, the dependency degree of condition attribute set can reflect the real condition better. This paper proposed a calculation method for dependency degree of condition attribute set based on discernibility matrix. The theoretical analysis and the simulation experiment results all show that the proposed method has better effectiveness and lower time complexity, which is valuable in attribute reduction in practice.

Acknowledgements

The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions, which have improved the presentation. This work was supported in part by the Science and Technology Plan Projects of Henan Province of China under grant No. 152102210357 and No. 152102210149, the Youth Backbone Teachers Funding Planning Project of Colleges and Universities in Henan Province of China under grant No.2014GGJS-084, the Key Science Research Project of Colleges and Universities in Henan Province of China under grant No. 16A520030, the Youth Backbone Teachers Training Targets Funded Project of Zhengzhou University of Light Industry of Henan Province of China under grant No.XGGJS02, the Ph.D. Research Funded Project of Zhengzhou University of Light Industry of Henan Province of China under grant No.2010BSJJ038 and No.2014BSJJ080, and the National Science Foundation of China under grant No.81501548.

References

1. H. Dan, X. C. Yu, "Statistical Inference of Rough Set Dependence and Importance Analysis," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 6, pp. 1070-1079, 2013.
2. J. Fan, Y. L. Jiang, Y. Liu, "Quick attribute reduction with generalized indiscernibility models," *Information Sciences*, vol. 397, pp. 15-36, 2017.
3. S. T. Hu, Y. Q. He, "Rough decision theory and application," Beihang University Press, Beijing, 2006.
4. D. Hu, X. C. Yu, J.Y. Wang, "Statistical Inference in Rough Set Theory Based on Kolmogorov–Smirnov Goodness-of-Fit Test," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 799-812, 2017.
5. Y. Y. Huang, T. R. Li, C. Luo, et al, "Dynamic variable precision rough set approach for probabilistic set-valued information systems," *Knowledge-Based Systems*, vol. 122, pp. 131-147, 2017.
6. U. Jamal, G. Rozaida, M. M. Deris, "An Empirical Analysis of Rough Set Categorical Clustering Techniques," *Plos One*, vol. 12, no. 1, pp. 1-22, 2017.
7. A. Joshuva, V. Sugumaran , "Classification of Various Wind Turbine Blade Faults through Vibration Signals Using Hyperpipes and Voting Feature Intervals Algorithm," *International Journal of Performability Engineering*, vol. 13, no. 3, pp. 247-258, 2017.
8. J. Konecny, "On attribute reduction in concept lattices: Methods based on discernibility matrix are outperformed by basic clarification and reduction," *Information sciences*, vol.415, pp.199-212, 2017.
9. J. Y. Liang, F. Wang, C. Y. Dang, et al "An efficient rough feature selection algorithm with a multi-granulation view," *International Journal of Approximate Reasoning*, vol. 53, pp.912-926, 2012.
10. L. Liu, B. S. Wang, Q. X. Zhong, et al, "A New Method for Decision Tree Based Discernibility Matrix and Degree of Consistent Dependence," *Applied Mechanics and Materials*, vol.743, pp. 390-394, 2015.
11. W.M. Ma, B. Z. Sun, "Probabilistic rough set over two universes and rough entropy," *International Journal of Approximate Reasoning*, vol. 53, no. 4, pp. 608-619, 2012.
12. C. U. Mba, H. A. Gabbar, S. Marchesiello, et al, "Fault Diagnosis in Flywheels: Case Study of a Reaction Wheel Dynamic System with Bearing Imperfections," *International Journal of Performability Engineering*, vol. 13, no. 4, pp. 362-373, 2017.
13. I. K. Park, G. S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289-295, 2015.
14. J. Qian, C. Y. Dang, X. D. Yue, et al, "Attribute reduction for sequential three-way decisions under dynamic granulation," *International Journal of Approximate Reasoning*, vol. 85, pp. 196-216, 2017.
15. M. Salamó, M. López-Sánchez, "Rough set based approaches to feature selection for case-based reasoning classifiers," *Pattern Recognition Letters*, vol. 32, no. 2, pp.280-292, 2011.
16. A. Sanchis, M. J. Segovia, J. A. Gil, et al, "Rough Sets and the role of the monetary policy in financial stability (macroeconomic problem) and the prediction of insolvency in insurance sector (microeconomic problem)," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1554-1573, 2007.
17. Y. H. She, X. L. He, H. X. Shi, et al, "A multiple-valued logic approach for multigranulation rough set model, " *International Journal of Approximate Reasoning*, vol. 82, no. 1, pp. 270-284, 2017.
18. B. Yang, B. Q. Hu, "On some types of fuzzy covering-based rough sets," Fuzzy *Sets and Systems*, vol. 312, pp. 36-65, 2017.
19. X. X. Zhang, D. G. Chen, E. C. C. Tsang, "Generalized dominance rough set models for the dominance intuitionistic fuzzy information systems," *Information Sciences*, vol. 378, pp. 1-25, 2017.
20. X. Y. Zhang, D. Q. Miao, "Three-way attribute reducts," *International Journal of Approximate Reasoning*, vol. 88, pp.401-434, 2017.

Hongchan Li was born in Hebei Province, China, in 1983. She received her B.S. degree from Heilongjiang Bayi Agricultural University, Daqing, Heilongjiang Province, China, in 2007 and her M.S. degree from Sichuan University of Science and Engineering, Zigong, Sichuan Province, China, in 2010. Since 2010, she has been with the faculty of the School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan Province, China, where she is currently a lecturer. Her major research interests include Cloud Computation, Intelligence Information Processing, Computing Intelligence and Data Mining.

Junxing Liu was born in Henan Province, China, in 1980. He received his M.S. degree from Central China Normal University, Wuhan, Hubei Province, China. Since 2006, he has been a chief coach of informatics Olympiad. In 2017, he had the honor of receiving the title of Zhengzhou City Outstanding Teacher. He is also a national gold medal coach. He has been counseling dozens of people who received the National Youth Olympic League first prize. He won the first prize of the national quality class competition in 2014. His major research interests include Information Technology Education, Program design and Smart Education.

Haodong Zhu was born in Henan Province, China, in 1980. He received his B.S. degree from Lanzhou Jiaotong University, Lanzhou, Gansu Province, China, in 2004, his M.S. degree from Sichuan University of Science & Engineering, Zigong, Sichuan Province, China, in 2008, and his Ph.D. degree from Graduate University of Chinese Academy of Sciences in 2011. Since 2010, he has been with the faculty of the School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan Province, China, where he is currently an Associate Professor and a master Tutor. His major research interests include Cloud Computation, Intelligence Information Processing, Computing Intelligence and Data Mining.