Two Algorithms for Finding Mutually Exclusive Features in Information Tables

Feng Honghai[1,2], Liu Baoyan[3], He Liyun[3],
Huang Yong[4], Yang Bingru[2], and Chen Yumei[5]

[1] Urban & Rural Construction School, Hebei Agricultural University, Postal Code 071001, Baoding, China
Honghf@hebau.edu.cn
[2] University of Science and Technology Beijing, Postal Code 100083, Beijing, China
[3] China Academy of Traditional Chinese Medicine, Postal Code 100700, Beijing, China
[4] Hebei Agricultural University, Postal Code 071001, Baoding, China
[5] Tian'e Chemical Fiber Company of Hebei Baoding, 071000 Baoding, China

Abstract. In real-world data sets we often find that some features' states are not true at the same time, which is a kind of very useful comparative knowledge for domain experts. For example, "coprostasis" and "diarrhea" are two mutually exclusive attributes. In this paper, we introduce two algorithms and find some mutually exclusive attributes from the SARS data set that have not been detected by medical experts in clinical practice, which are" infected by bacteria" and "damage of heart", "Infected by bacteria" and "pathological changes in the lung", "Infected by bacteria" and "damage of heart" etc.

1 Introduction

In a real-world data set, we often find that there are mutually exclusive correlations between two attributes, that is, the two attributes cannot be true in the same time, which is kind of very useful comparative knowledge for domain experts. As we know, "coprostasis" and "diarrhea" are two mutually exclusive attributes, which means that the states of coprostasis and diarrhea cannot be true together, i.e., a patient cannot suffer from coprostasis and diarrhea at the same time. Some machine learning theories can induce meaningful rules from data set. Association rules [1], support vector machines [2-4], neural networks [5-7] and decision trees [8] can extract rules from data sets, but they have not been used to induce this kind of knowledge.

Rough Sets theory [9] developed by Zdzislaw Pawlak in the early 1980's can search large databases for meaningful decision rules and finally acquire new knowledge.

In this paper, using some rough set concepts we propose a new concept of probability equivalence between two sets and based on it we present two algorithms for detecting the mutually exclusive features in information tables.

2 Rough Set Basic Concepts

2.1 Basic Concepts of the Rough Sets Theory

Knowledge about objects is often represented in the form of an information table, the rows of the table are labeled by objects, columns are labeled by attributes and entries of the table are attribute-values, called descriptors. Formally, by an information table we understand a 4-tuple $S = \langle U, Q, V, f \rangle$, where U is a finite set of objects, Q is a finite set of attributes, $V = \cup_{q \in Q} V_q$, where V_q is the domain of attribute q , and $f : U \times Q \to V$ is a total function such that $f(x, q) \in V_q$ for every $(x, q) \in U \times Q$, called an information function .The set Q is, in general, divided into the set C of condition attributes and the set D of decision attributes.

2.2 Indiscernibility Relation

For every set of attributes $B \subset A$, an indiscernibility relation $IND(B)$ is defined in the following way: two objects, x_i and x_j , are indiscernible by the set of attributes B in A , if b (x_i)=b (x_j) for every b $\subset B$. The equivalence class of $IND(B)$ is called elementary set in B because it presents the smallest discernible groups of objects. For any element x_i of U , the equivalence class of x_i in relation $IND(B)$ is represented as $[x_i]_{IND(B)}$. The construction of elementary sets is the first step in classification with rough sets.

Let $F = \{X_1, X_2, \cdots, X_n\}$, $X_i \subset U$ be a family of subsets of the universe *U*. If the subsets in *F* do not overlap, i.e., $X_i \bigcap X_j \neq \Phi$ and the entity of them contains all elementary sets, i.e., $\bigcup X_i = U$ for $i = 1, \cdots, n$, then, *F* is called a classification of *U*, whereas *X* are called classes.

2.3 Some Measures to Describe Dependency between Attributes

The measures to describe inexactness of approximate classifications have been defined; the first measure is the accuracy of approximation of F by B . It expresses the possible correct decisions when classifying objects employing the attribute B .

$$d_B(F) = \sum_{i=1}^{n} \left|B_-(X)\right| / \sum_{i=1}^{n} \left|B^-(X)\right|$$

The second measure is called the quality of approximation of F by B. It expresses the possible correct decisions when classifying objects employing the attribute B.

$$r_B(F) = \sum_{i=1}^{n} \left|B_-(X_i)\right| / \left|U\right|$$

The third measure is Mutual information. Mutual information represents a general information theoretic approach to determine the statistical dependence between variables (attributes). The concept was initially developed for discrete data. For a system, *A*, with a finite set of *M* possible states $\{a_1, a_2, a_{M_A}\}$, the Shannon entropy *H*(*A*) is defined as [10]

$$H(A) = -\sum_{i=1}^{M_A} p(a_i) \log p(a_i)$$

where $p(a_i)$ denotes the probability of the state a_i. The Shannon entropy is a measure for how evenly the states of *A* are distributed. The entropy of system *A* becomes zero if the outcome of a measurement of *A* is completely determined to be a_j, thus if $p(a_j) = 1$ and $p(a_i) = 0$ for all $i \neq j$, whereas the entropy becomes maximal if all probabilities are equal. The joint entropy *H* (*A, B*) of two systems *A* and *B* is defined analogously

$$H(A,B) = -\sum_{i=1,j=1}^{M_A,M_B} p(a_i,b_j) \log p(a_i,b_j)$$

This leads to the relation

$$H(A, B) \leq H(A) + H(B)$$

which fulfils equality only in the case of statistical independence of *A* and *B*. Mutual information *MI*(*A, B*) can be defined as [11]

$$MI(A, B) = H(A) + H(B) - H(A, B) \geq 0$$

It is zero if *A* and *B* are statistically independent and increases the less statistically independent *A* and *B* are.

3 Algorithm for Inducing the Mutually Exclusive Property between Two Attributes

3.1 Probability Equivalence between Sets

The cardinality of the intersection of condition attribute equivalence classes and decision attribute equivalence classes embodies the fact that the examples classified with condition attribute is according to the examples classified with decision attribute.

In conventional rough set methods, the induced rules are of the form " $A \rightarrow B$ ", not the" $A \leftarrow B$ " and " $A \leftrightarrow B$ ". Due to $A \leftrightarrow B$ $\Leftrightarrow$ ($A \rightarrow B$) $\wedge$ ($A \leftarrow B$), $A \leftrightarrow B$ embodies the equivalence of the two sets, $\frac{|A \cap B|}{|A|}$ embodies the sufficiency of " $A \rightarrow B$ ", $\frac{|A \cap B|}{|B|}$ embodies the necessity of " $A \rightarrow B$ ", $\frac{|A \cap B|}{|A|} \bullet \frac{|A \cap B|}{|B|}$ embodies the probability equivalence of " $A \leftrightarrow B$ ", namely the joint probability of " $A \rightarrow B$ " and " $A \leftarrow B$ ".

The degree of probability equivalence of $A \leftrightarrow B$ is defined as

$$s = \frac{|A \cap B|}{|A|} \bullet \frac{|A \cap B|}{|B|} \ (\%)$$

3.2 Discussions about the Measure of Probability Equivalence

(1) If $A = B$, then $s = 1$; if $|A \cap B| = 0$, then $s = 0$; if $|A \cap B| \neq 0$ and $A \neq B$, then $0 < s < 1$. So we can conclude that $0 \leq s \leq 1$.

(2) If $|A \cap B| \neq 0$ and $A \neq B$, $\frac{|A \cap B|}{|A|}$ is the value of rough membership function with regard to $A \rightarrow B$, $\frac{|A \cap B|}{|B|}$ is the value of rough membership function with regard to $A \leftarrow B$.

(3) The measure of probability equivalence holds the properties such as non-negative, reflexivity and transitivity.

Proof:

Given a decision table $T = (U, C \bigcup D)$, C is a condition attribute set and D a decision attribute, respectively, $A \in C$, $B \in C$ are two condition attributes. d denotes the distance between two attributes.

Obviously, non-negative and reflexivity can be held. Now we give the proof of the transitivity.

(1) When A and B are the same entirely, $S(B|A) = S(A|B) = 1$, $d_{A \rightarrow B} = 0$, $d_{A \rightarrow B} + d_{B \rightarrow D} = d_{A \rightarrow D} = d_{B \rightarrow D}$

(2) When the intersections of equivalence classes of A and B are empty, $S(B|A)=S(A|B)=0$, and $d_{A\to B}=1$. Due to $MAXd_{A\to D}=1$, and A cannot be the same as D, then we can hold that $d_{A\to D}<d_{A\to B}+d_{B\to D}=1+d_{B\to D}$.

Because A and B cannot be the same, during the course of A changing from being the same as B to being not entirely the same as B, $S(B|A)$ is changing from the largest to the least (from 1 to 0), and $d_{A\to B}$ from the least to the largest monotonously, we can hold that $d_{A\to D}<d_{A\to B}+d_{B\to D}$, so the property of transitivity is held.

(3) In conclusion, the measure of probability equivalence embodies the equivalence of two sets.

3.3 Algorithms

According to the error theory, the error proportion caused by mistakes in repetitive experiments approximated to 0.3%, but in the real world data set the mistake would be as much as 5%. So based on the above two results the two algorithms are presented as follows.

Algorithm 1:

(1) Select any two attributes C_i and C_j that are Boolean variables, i.e., $V_{C_i}=V_{C_l}=\{0,1\}$.

(2) For attributes C_i and C_j, calculate the equivalence class C_{i1} with $V_{C_i}=1$ and C_{j1} with $V_{C_j}=1$.

(3) Calculate the value of s of the two equivalence classes C_{i1} and C_{j1}.

(4) Take the threshold of $s=e\times e$ =0.3 × 0.3%=0.09%, which is because $A\leftrightarrow B \Leftrightarrow (A\to B)\wedge(A\leftarrow B)$. Whereas for $(A\to B)\vee(A\leftarrow B)$, we should take the threshold of s = 0.3%+0.3%=0.6% contrarily, which does not accord with our case. If the value of s is less than 0.09%, we regard that the values in the intersection- $A\bigcap B$ of attribute *A* and *B* are mistakes, and regard that the attributes C_i and C_j are mutually exclusive, namely, as Boolean variables, attributes C_i and C_j do not take the value 1 simultaneous, or are not true at the same time.

Algorithm 2:

The difference between algorithm 2 and algorithm 1 is step 2, that is, in algorithm 2 we use "SORT" command of SQL to sort the selected two attributes C_i and C_j in descending order of attribute values and to calculate $A\bigcap B=(IND\ (C_i{=}1)\cap IND\ (C_j{=}1))$ and the value of s. So man-machine interaction is the advantage of using algorithm 2.

4 An Illustrative Example

In Table 1, U= {1, 2, 3, 4, 5, 6, 7}, condition attribute set C= {a, b, c, d}, decision attribute set D= {e}, a, b, c, d are Boolean variables. $IND(a=1)$= {1,2,4,5}, $IND(b=1)$= {3,6}, $IND(c=1)$= {3,4}, $IND(d=1)$= {1,4,5}, only $IND(a=1) \cap IND(b=1)=\Phi$, so we can conclude that c and d are mutually exclusive attributes.

Table 1 An information table

U	a	b	c	d	e
1	1	0	0	1	1
2	1	0	0	0	1
3	0	1	1	0	0
4	1	0	1	1	0
5	1	0	0	1	2
6	0	1	0	0	2
7	0	0	0	0	2

5 UCI-Votes Data Set Experiments

In the votes data set there are 17 attributes that are “Class Name”, “physician-fee-freeze”, “export-administration-act-south-africa”, “duty-free-exports”, “crime”, “superfund-right-to-sue”, “education-spending”, “synfuels-corporation-cutback”, “immigration”, “mx-missile”, “aid-to-nicaraguan-contras”, “el-salvador-aid”, “anti-satellite-test-ban”, “religious-groups-in-schools”, “water-project-cost-sharing”, “adoption-of-the-budget-resolution”, and “handicapped-infants”, and there are 434 examples in all. Using the above algorithms we find that attributes “aid-to-nicaraguan-contras” and “el-salvador-aid” are mutually exclusive attributes. Among the total 434 examples there are 242 whose values of “aid-to-nicaraguan-contras” are “yes” and there are 211 whose values of “el-salvador-aid” are “yes”, additionally, there are only 31 examples whose values of “aid-to-nicaraguan-contras” and “el-salvador-aid” are “yes” at the same time. So $s= (31\times31)/ (242\times211) =0.019$.

6 SARS Data Set Experiments

Usually we have the knowledge of correlation between two attributes, but sometimes we may have little knowledge about them.

The SARS data set is acquired from Beijing. Using the method mentioned above we obtain some rules that embody mutually exclusive property. The following are part of mutually exclusive attributes some of which are not observed by medical experts in clinical practice.

(1) "Constipation" and "diarrhea", s= (2×2)/ (251×273) =0.0006. Obviously, the two cases are mistakes, i.e., the two patients who suffer from constipation and diarrhea at the same time are mistake cases.

(2) "ARDS" and "MODS" (Acute Respiratory Distress Syndrome), s= 0/ (135×37) =0, which means that there are 135 patients who suffer from ARDS, and there are 37 patients who suffer from MODS. Furthermore, there is no SARS patient who suffers from MODS and ARDS at the same time.

(3) "MODS" and "damage of heart", s=1/ (37×611) =0.000044, which means that there are 37 patients who suffer from MODS, and there are 611 patients whose hearts are damaged. Furthermore, there is only one SARS patient who suffers from MODS and whose heart is damaged at the same time. Maybe this case is a mistake.

(4) "MODS" and "damage of function of liver", s=0/ (37×1803) =0, which means that there are 37 patients who suffer from MODS, and there are 1803 patients whose livers are damaged. Furthermore, there is no SARS patient who suffers from MODS and whose liver is damaged at the same time.

(5) "Infected by fungi" and "MODS", s=0/ (105×37) =0, which means that there are 105 patients who are infected by fungi, and there are 37 patients who suffer from MODS. Furthermore, there is no SARS patient who is infected by fungi and suffers from MODS at the same time.

(6) "Infected by bacteria" and "damage of heart", s=0/ (589×611) =0, which means that there are 589 patients who are infected by bacteria, and there are 611 patients whose hearts are damaged. Furthermore, there is no SARS patient who is infected by bacteria and whose heart is damaged at the same time.

(7) "Infected by bacteria" and "moist rale", s=0/ (589×139) =0, which means that there are 589 patients who are infected by bacteria, and there are 139 patients who suffer from moist rale. Furthermore, there is no SARS patient who is infected by bacteria and suffers from moist rale at the same time.

(8) "Infected by bacteria" and "rhohchi", s=0/ (589×76) =0, which means that there are 589 patients who are infected by bacteria, and there are 76 patients who suffer from rhohchi. Furthermore, there is no SARS patient who is infected by bacteria and suffers from rhohchi at the same time.

(9) "Infected by bacteria" and "pathological changes in the lung", s=0/ (589×114) =0, which means that there are 589 patients who are infected by bacteria, and there are 114 patients who suffer from pathological changes in the lung. Furthermore, there is no SARS patient who is infected by bacteria and suffers from pathological changes in the lung at the same time.

7 Conclusions and Discussions

(1) Since in algorithm 2 the "SORT" command of SQL is to be used to sort the selected two attributes in descending order of attribute values to calculate $A \cap B$ and the value of s, we can see the IND (C_i=1), IND (C_j=1) and $A \cap B$ =(IND (C_i=1) ∩ IND (C_j=1))clearly in the data set. So the algorithm 2 is friendly to the user.

(2) The definition of probability equivalence can measure the degree of causal correlation between two equivalence classes such as $A \leftrightarrow B$, whereas the accuracy of approximation and the quality of approximation only measure the dependency of one attribute C_i on another attribute C_j ($C_i \rightarrow C_j$). Because mutually exclusive attributes embody the correlation between two equivalence classes, the measures that only embody the dependency of one equivalence class on another equivalence class such as $A \rightarrow B$ cannot be used to generate the mutually exclusive attributes.

(3) The measure of probability equivalence embodies the distance between two sets (equivalence classes), i.e., it holds the properties such as non-negative, reflexivity and transitivity.

(4) For mutually exclusive attributes, their values cannot take 1 (true) at the same time, but their values can take 0 (false) at the same time, and at the time when one attribute's values take 0 (or 1), the other attribute's values can take 1 (or 0).

(5) For the measures of accuracy of approximation and quality of approximation, they all measure dependency of a decision attribute on a condition attribute, in other words, they do not embody the correlation between two equivalence classes generated with the respective attributes.

(6) For the measure of mutual information, it embodies the correlation between two attributes whereas it does not embody the correlation between two equivalence classes of the two corresponding attributes.

(7) Association rules measure the degree of items' (attributes') state being true at the same time, for example, in a supermarket, milk and diapers are sold at the same time. So association rules do not measure the degree of items' state not being true at the same time, namely, it does not measure the mutually exclusive property between two attributes.

Acknowledgements

The paper is supported by the Special Foundation of SARS of the National High-Tech Research and Development Major Program of China (863)- "Clinical Research on the Treatment of Severe Acute Respiratory Syndrome with Integrated Traditional and Western (No.2003AA208101)"

References

[1] J.Grzymala-Busse: *Managing Uncertainty in Expert Systems.* Kluwer Academic Publisher, Boston, 1991

[2] J. Diederich , and N. Barakat, "Hybrid rule-extraction from support vector machines" in Proc. of IEEE conference on cybernetics and intelligent systems, Singapore, 2004, pp. 1270-1275.

[3] Andreu Catal Haydemar Nez, Cecilio Angulo. "Rule extraction from support vector machines". In ESANN'2002 proceedings - European Symposium on Artificial Neural Networks, pp. 107–112. Dside, 2002.

[4] H. Núñez, C. Angulo, and A.Catala, "Rule-extraction from Support Vector Machines", in Proc. of European Symposium on Artificial Neural Networks, Burges, 2002, pp.107-112.
[5] A. Tickle R. Andrews, J. Diederich. "A survey and critique of techniques for extracting rules from trained artificial neural networks". Knowledge-Based Systems, Vol. 8, pp. 373–389, 1995.
[6] G. Towell & J. Shavlik. "The extraction of refined rules from knowledge-based neural networks". Machine Learning, Vol. 13, pp. 71–101, 1993.
[7] F. J. Kurfes. Neural networks and structured knowledge: Rule extraction and applications. Applied Intelligence (Special Issue), Vol. 12(1-2), pp. 7–13, 2000.
[8] S. G. Fountoukis, M. P. Bekakos, J. P. Kontos. "Rule extraction from decision trees with complex nominal data", Neural, Parallel & Scientific Computations, Vol. 9(2) pp. 119-128, 2001.
[9] Z. Pawlak: Rough Sets: *Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publisher, Norwell, MA, 1999.
[10] Jonathan, I., Maletic Andrian Marcus, "Data Cleaning: Beyond Integrity Analysis". Division of Computer Science, Vol. 2, 2000
[11] Shannon CE, "A mathematical theory of communication". The Bell System Technical Journal, Vol. 27, pp. 623-656, 1948

Biography

▲ Name: Feng Honghai
Address: Urban & Rural Construction School, Hebei Agricultural University, 071001, Baoding, China
Education & Work experience: He is persuing a doctor degree in University of Science and Technology Bei-jing
Tel: +86-312-7521102
E-mail: honghf@hebau.edu.cn