Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Rough Set Based Classification rules generation for SARS Patients

Feng Honghai, Chen Guoshun, Wang Yufeng, Yang Bingru, Chen Yumei

***Abstract*—SARS is an acute infectious disease and can cause a large amount of death. Up until now we have not known it well. With the experimental results of micronutrients of 30 SARS patients and 30 non-SARS patients, using rough set theory we induce some classification rules. Attribute reduction results show that micronutrients Fe, Ca, K and Na are necessary and sufficient for classification, whereas micronutrients Zn, Cu and Mg are not necessary or are redundant. Additionally, we find that micronutrient Ca has a strong correlation to SARS. The classification results of 30 other examples show that the rough set classification method is available.**

I. Introduction

SARS is an acute infectious disease and can cause a large amount of death. The diagnosis of SARS is still an important issue for doctors and researchers. Up until now we have not known it well. Due to the little experience of the diagnosis of SARS, little knowledge can be used to guide future clinical diagnosis. So knowledge acquisition is the main approach to resolve the problem.

A method for knowledge acquisition is automated discovery from observed or tested data, that is, to design an algorithm which can learn and refine decision rules from a set of training samples, or observed data. This method is referred to as data mining (DM) or knowledge discovery (KDD) [1].

In order to automate this problem, many inductive learning methods, such as induction of decision trees [2,3], rule induction methods [4,5], association rule [6] and rough set theory [7,8,9], are introduced and applied to extract knowledge from databases, and the results show that these methods are appropriate.

Some new KDD or DM techniques such as artificial neural networks (ANN) and support vector machine (SVM) etc are black-box ones, i.e., they cannot be used to induce the understood knowledge [10].

Rough set is a new knowledge discovery theory. Rough set can generate easily understood knowledge without additional information outside of the data set.

This paper presents an application of a rough set approach for the automated discovery of classification rules form a data set that is the experimental result of micronutrients of 30 SARS patients and 30 non-SARS patients. Using rough set theory we induce some classification rules. Attribute reduction results show that micronutrients Fe, Ca, K and Na are necessary and sufficient for classification, and micronutrients Zn, Cu and Mg are not necessary or are redundant. Additionally, we find that micronutrient Ca has a strong correlation to SARS. Other 30 test examples' classification results show that the rough set classification method is available.

Manuscript received May 2, 2005.

Feng Honghai is with the Hebai Agriculture University and the University of Science and Technology Beijing, Beijing 100083 P.R. China (Home phone: (86-10) 62391821; e-mail: honghf@ mail.hebau.edu.cn).

Chen Guoshun is with the Ordnance Technology Institute, 050000 Shijiazhuang, China.

Wang Yufeng is with the University of Science and Technology Beijing, Beijing 100083 P.R. China.

Yang Bingru is with the University of Science and Technology Beijing, Beijing 100083 P.R. China.

Chen Yumei is with Tian'e Chemical Fiber Company of Hebei Baoding, 071000 Baoding, China

II. Rough Set Concepts

A. Decision Table

Data are often presented as a table, columns of which are labeled by attributes, rows by objects of interest and entries of the table are attribute values. Such tables are known as information systems, decision tables, or information tables.

A decision table is composed of a 4-tuple DT=$<U, A, V, f>$, where $U=\{x_1, x_2, \ldots n\}$ is a nonempty, finite set called the universe; A is a nonempty, finite set of attributes; $A=C\cup D$, in which C is a finite set of condition attributes and D is a finite set of decision attributes; $V=\bigcup_{a\in A} V_a$, where V_a is a domain (value) of the attribute a, and $f: U\times A\to V$ is called the information function such that $f(x,a)\in V_a$ for every $a\in A, x_i\in U$.

B. Indiscernibility Relation

In a decision table DT=$<U, A, V, f>$, to every subset of attributes $B\subseteq A$, a binary relation, denoted by $IND(B)$, called the B-indiscernibility relation, is associated and defined as follows:

$$IND(B)=\{(x_i, x_j)\in U\times U : a\in B, f(x_i,a)=f(x_j,a)\} \quad (1)$$

C. Reduction

In a decision table DT=$<U, A, V, f>$, the reduction of condition attribute C means a nonempty subset $R\subset C$ satisfied by the following condition:

1) $IND(R)=IND(C)$

2) There is no subset of $R' \subset R$, which is satisfied by $IND(R') = IND(C)$

The reduced set of attributes provides the same ability of classification as the original set of attributes.

D. Discernibility Matrix

By an discernibility matrix of $B \subseteq A$ denoted $M(B)$ we will mean $n \times n$ matrix defined as:

$$c_{ij} = \begin{cases} a \in C : f(x_i) \neq f(x_j, a) \wedge (d \in D, f(x_j, d) \neq f(x_i, d)) \\ 0, d \in D, f(x_i, d) = f(x_j, d) \\ \phi, f(x_i, a) = f(x_j, d) \wedge (d \in D, f(x_i, d) \neq f(x_j, d)) \end{cases} \quad (2)$$

Where $i, j = 1, 2, \ldots n$

Thus entry c_{ij} is the set of all attributes that classify objects x_i and x_j into different decision classes in U. From formula (2), all of the distinguishing information for attributes is contained in above discernibility matrix.

It is easily seen that the core is the set of all single element entries of the discernibility matrix $M(B)$, i.e.,

$$CORE(B) = \{a \in B : c_{ij} = \{a\}, \text{ for some } i, j\}$$

Obviously $B' \subseteq B$ is a reduct of B, if B' is the minimal (with respect to inclusion) subset of B such that

$B' \cap c \neq \varnothing$ for any nonempty entry $c(c \neq \varnothing)$ in $M(B)$.

III. SARS Data Set Experiments

A. SARS Data Set

Table I gives partial experimental results of micronutrients that are essential in minute amounts for the proper growth and metabolism of human beings. Among them, examples 31~60 are the results of SARS patients and 61~90 are the results of non-SARS patients. In Table I, 31~37 are non-SARS patients, and 61~67 are SARS patients. Attribute "1" denotes micronutrient Zn, attribute "2" denotes micronutrient Cu, attribute "3" denotes micronutrient Fe, attribute "4" denotes micronutrient Ca, attribute "5" denotes micronutrient Mg; attribute "6" denotes micronutrient K, attribute "7" denotes micronutrient Na, and decision attribute "C" denotes the class "SARS" and "no SARS". $V_C = \{0, 1\}$, where "0" denotes "no SARS", "1" denotes "SARS".

TABLE I PARTIAL EXPERIMENT RESULTS OF MICRONUTRIENTS OF HUMAN BEINGS

U	1	2	3	4	5	6	7	C
31	166	15.8	24.5	700	112	179	513	0
32	185	15.7	31.5	701	125	184	427	0
33	193	9.80	25.9	541	163	128	642	0
34	159	14.2	39.7	896	99.2	239	726	0
35	226	16.2	23.8	606	152	70.3	218	0
36	171	9.29	9.29	307	187	45.5	257	0
37	201	13.3	26.6	551	101	49.4	141	0
61	213	19.1	36.2	2220	249	40.0	168	1
62	170	13.9	29.8	1285	226	47.9	330	1
63	162	13.2	19.8	1521	166	36.2	133	1
64	203	13.0	90.8	1544	162	98.9	394	1
65	167	13.1	14.1	2278	212	46.3	134	1
66	164	12.9	18.6	2993	197	36.3	94.5	1
67	167	15.0	27.0	2056	260	64.6	237	1

B. Rough Set Based Classification

(1) Discretization

Table II is the result of discretization and attribute reduction. For attribute "3" (Fe), 0 denotes [4.04, 8.65], 1 denotes [8.65, 53.3], 2 denotes [53.3, 322]. For attribute "4" (Ca), 0 denotes [135, 1025], 1 denotes [1025, 1437], 2 denotes [1437, 6747]. For attribute "6" (K), 0 denotes [31, 45.2], 1 denotes [45.2, 228], 2 denotes [228, 1314]. For attribute "7" (Na), 0 denotes [67.3, 126], 1 denotes [126, 899], 2 denotes [899, 1372].

After discretization, some examples become a repeat: among which 32, 33, 35 and 37 are repeats; 31, 36, 38, 40, 44, 45 and 46 are repeats; 47 and 57 are repeats; 50 and 51 are repeats; 53 and 62 are repeats; 56 and 58 are repeats; 65, 89 and 90 are repeats; 67, 77, 83 and 84 are repeats; 72 and 82 are repeats; 74, 75, 80 and 81 are repeats respectively.

(2) Attribute reduction

If the discernibility matrix is used to execute attribute reduction and value reduction directly, the complexity is higher. Since the amount of condition attributes are only 7, and the amount of examples are only 19 after eliminating the repeated ones, we use another attribute reduction method that is described as follows:

After eliminating attributes 1, 2 and 5 and their values respectively, the examples left are not inconsistent, so attributes 1,2 and 5 are redundant attributes. While after eliminating attributes 3, 4, 6, 7 and their values respectively, the examples left become inconsistent ones, so attributes 3, 4, 6, and 7 are core attributes. Furthermore, attributes 3, 4, 6, and 7 are sufficient for describing the information system. In other words, after eliminating attributes 1, 2, and 5, the examples left are consistent, while if eliminating one of the attributes from 3, 4, 6, and 7, the information table will become inconsistent.

Table II describes the results after discretization and attribute reduction. Table III describes the results of discerniblity matrix

TABLE II RESULTS AFTER DISCRETIZATION AND ATTRIBUTE REDUCTION

U	3	4	6	7	C
31	1	0	1	1	0
34	1	0	2	1	0
39	0	0	1	1	0
41	1	0	2	2	0
48	0	2	1	1	0
49	1	1	1	2	0
52	2	0	1	1	0

53	1	1	0	1	0
59	0	0	2	1	0
61	1	2	0	1	1
62	1	1	1	1	1
64	2	2	1	1	1
65	1	2	1	1	1
66	1	2	0	0	1
68	1	1	1	0	1
69	1	2	1	2	1
70	2	2	2	1	1
71	2	1	1	1	1
86	1	2	1	0	1

TABLE III DISCERNIBLITY MATRIX

	31	34	39	41	48	49	52	53	59
61	4,6	4,6	A	B	3,6	B	A	4	A
62	4,6	4,6	3,4	B	3,4	7	3,4	6	A
64	3,4	A	3,4	C	3	D	4	A	A
65	4	4,6	3,4	B	3	4,7	3,4	4,6	A
66	B	B	C	B	3,6,7	C	C	4,7	C
68	4,7	B	D	B	D	7	D	6,7	C
69	4,7	B	D	4,6	3,7	4	D	B	C
70	A	3,4	A	D	3,6	C	4,6	A	3,4
71	3,4	A	3,4	C	3,4	3,7	4,6	3,6	A
86	4,7	B	D	B	3,7	4,7	D	B	C

where A denotes "3,4,6"; B denotes "4,6,7"; D denotes "3,4,7"; C denotes "3,4,6,7" respectively.

(3) Value reduction and rule generation

Each row (example) of a decision table determines a decision rule, which specifies decisions (actions) that should be taken when conditions pointed out by condition attributes are satisfied.

Using the concept of a value reduct, Table II and Table III can be simplified as follows (Table IV).

TABLE IV VALUE REDUCTION AND RULE GENERATION

U	3 (Fe)	4 (Ca)	6 (K)	7 (Na)	C (SARS)
31,34,39,41,52,59		0			0
48	0				0
49		1		2	0
53		1	0		0
61,65,69	1	2			1
62		1	1	1	1
64,70	2	2			1
66,68,86				0	1
71	2	1			1

Table IV gives the most simplified classification rules set. From Table IVwe get the following classification rules:

(1) Ca=0 → C=0
(2) Fe=0 → C=0
(3) Ca=1 and Na=2 → C=0
(4) Ca=1 and K=0 → C=0
(5) Fe=1 and Ca=2 → C=1
(6) Ca=1 and K=1 and Na=1 → C=1
(7) Fe=2 and Ca=2 → C=1
(8) Na=0 → C=1
(9) Fe=2 and Ca=1 → C=1

C. Classification of Some SARS Patients and Healthy Human Beings

Table V gives the classification results of 14 examples that include SARS patients and healthy human beings. The classification precision is 100%.

TABLE V PARTIAL EXPERIMENT RESULTS OF MICRONUTRIENTS OF HUMAN BEINGS

U	1	2	3	4	5	6	7	C
1	58.2	5.42	29.7	323	138	179	513	0
2	106	1.87	40.5	542	177	184	427	0
3	152	0.80	12.5	1332	176	128	646	1
4	85.5	1.70	3.99	503	62.3	238	762	0
5	144	0.70	15.1	547	79.7	71.0	218	0
6	85.7	1.09	4.2	790	170	45.8	257	0
7	144	0.30	9.11	417	552	49.5	141	0
8	170	4.16	9.32	943	260	155	680	0
9	176	0.57	7.3	318	133	99.4	318	0
10	192	7.06	32.9	1969	343	103	553	1
11	188	8.28	22.6	1208	231	1314	1372	1
12	153	5.87	34.8	328	163	264	672	0
13	143	2.84	15.7	265	123	73.0	347	0
14	213	19.1	36.2	2220	249	62.0	465	1

With rule (5) "Fe=1 and Ca=2 → C=1", examples 3, 10, 11 and 14 are classified into class 1. By rule (1) "Ca=0 → C=0", the other examples are classified into class 0.

IV. DISCUSSIONS

(1) Obviously, attribute Ca is the most important attribute. It has the strongest correlation to SARS or non-SARS. The attribute Fe takes the second place.

(2) The most simplified classification rules set should include the most important condition attribute as soon as possible.

(3) The rules "Ca=0 → C=0" and "Fe=0 → C=0 and Na=0 → C=1" imply that as long as the attribute Ca, Fe and Na take the value 0 classification knowledge can be obtained.

(4) If there is not the example 48 with $V_{Ca} = 2$, we can conclude that Ca=2 → C=1. Maybe this is an error or mistake. The following works are examining the mistake or error and add new test examples with $V_{Ca} = 2$.

REFERENCES

[1] B. G. Buchnan, E. H.Shortliffe. "Rule-based expert systems." New York: Addison-Wesley, 1984. .

[2] L. Breiman, J. Freidman, R. Olshen & C. Stone. Classification and regression trees. Belmont: Wadsworth International Group, 1984.

[3] J. R. Quinlan. C4.5-programs for machine learning. Palo Alto: Morgan Kaufmann, 1993.

[4] R. S. Michalski, I. Mozetic, J. Hong & N. Lavrac. The multipurpose incremental learning system aq15 and its testing application to three medical domains. Proceedings of the Fifth National Conference on Artificial Intelligence, 1986, pp. 1041–1045, Menlo Park: AAAI Press.

[5] J. W.Shavlik & Dietterich, T. G. (Eds.). Readings in machine learning. Palo Alto: Morgan Kaufmann, 1990.
[6] R.Agrawal, T.Imielinski & A Swami. Mining association rules between sets of items in large databases. Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93), pp.207–216.
[7] Z.Pawlak. Rough sets. Dordrecht: Kluwer Academic Publishers, 1991. .
[8] S.Tsumoto & H.Tanaka. PRIMEROSE: probabilistic rule induction method based on rough sets and resampling methods. Computational Intelligence, 1995, vol. 11, 389–405.
[9] W.Ziarko. Variable precision rough set model. Journal of Computer and System Sciences, 1993, vol. 46, 39–59.
[10] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy,"Improvements to the SMO algorithm for SVM regression," IEEE Trans. Neural Networks, vol. 11, pp. 1188–1193, Nov. 2000.