Rough Set-Based Prediction Approach to Medical Application

Jia-Yuarn Guo[1], ¡ fChun Pei[2], Yu-Tung Dai[3], Han Hsin Chou[4], Chung-hsiu Kuo[5]
[1]Dept. of Industrial Engineering and Management,
[2]The Graduate School of Gerontic Technology and Service Management, Nan Kai Institute of Technology
[3]Dept. of Occupational Safety and Health, Chung Haw College of Medical
[4]Department of Logistics and Shipping Management, Kainan University
[5]Ordnance Readiness and Development Center
Correspondence: Chun Pei, pantonpei@hotmail.com

Introduction

It is noted that due to the typically huge size of today's information systems, real-world data tend to be incomplete due to missing some values. Hence, discover knowledge from incomplete information systems has received more and more attention in recent years.

The rough sets theory, proposed by Pawlak , provides a natural method to cope with incomplete or inconsistent information which has been the mainly impediment to the classification and rule induction of objects. Several techniques have been developed to extract decision rules from an incomplete information system. A key factor among them is using different methods to manage the missing data (unknown values) . The simplest is removing the objects with unknown values. Other simple methods include replacing missing values with possible values calculated by statistics analysis. These methods all try to transform an incomplete systems into a complete system by smoothing or extending the data. Other groups of techniques deal with the incomplete systems without changing the size of the data sets or making assumption of the missing values. These methods intend to induce every certain rule directly from the original data sets. Like the second group, our approach uses a rule generation algorithm to induce all certain rules and possible rules from the original incomplete data. No matter what the missing values will be, they won't affect the induction rules.

Compared with most of the classification and rule induction methods which induce knowledge from sets approximation concepts, the rule induction technique proposed in this paper applies a new modified rule-reduct generation algorithm (MRGI) and rule induction program (RIPI) to mine knowledge directly from the minimal set of decision rules(rule-reduct).

Background and Definitions

A. Rough sets theory preliminary (R.S.T)

The rough sets theory provides a natural method to deal with incomplete or inconsistent information which has been the mainly obstacle to the classification and rule induction of objects.

In this section we recall some basic notation of R.S.T that related to our research, and we assume that reader is familiar with basic principle of R.S.T. For more detail introduction of R.S.T

-Incomplete information systems

An incomplete information system IS with two-tuple can be seen as a system: IS = (U, $\overline{A}$), where $\overline{A}$ is the set of attributes (features, variables) containing unknown values. Each attribute a $\in \overline{A}$ defines an information function : U ¡ $\overline{V_a}$, where $\overline{V_a}$ is the set of values of a including unknown values (M), called the domain of attribute a. A decision table is any information system with a decision attribute d: $\overline{T}$ =(U, $\overline{A}$ $\cup$ d).

Example.1 Transfer the incomplete medical diagnosis data (table1) into rough sets decision table.

Using the terminology of the rough sets theory this data set can be considered as follows:

U = {x0 , x1 , x2, x3, x4, x5, x6,. ….x9}
= {F1, F2, F3}
=(Temperature, Dry-cough , Headache)
The domains of the particular attributes are:
={0,1,2,M},
0=normal, 1= Subfebrile, 2=high, M=missing.
={0,1,M}, 0=absent, 1=present, M=missing.
={0,1,M}, 0=absent, 1=present, M=missing.
={0,1}, 0=absent, 1=present.

i.e., the domain of each attribute is the set of values of this attribute. The decision table for this system is presented in Table 2.

- Indiscernibility relation

For every set of input attributes R A, two objects, xi and xj have an indiscernibility relation Ind(B) when they are indiscernible by the set of attributes R in A, and if a(xi) = a(xj) for every a R. The equivalence class of Ind(R) or U/R ={E1 E2 ¡ KEm} is called elementary set in R because it represents the smallest discernible groups of objects.

As shown in Table2, there are some identical objects in the original data set, which cannot be distinguished based on input attributes in A. After grouping all objects based on three input/features (a1, a2 and a3) with the same values to remove indiscernibility, the results are shown in Table3.

B. Reduct generation

In R.S.T based applications for classification and rule induction, the knowledge induction from reducts is the most important concept. By definition, a reduct is defined as minimal sufficient sets of features necessary for the description of all features A,(Pawlak, 1991)[14]. Nevertheless , a rule reduct, r-reduct, is subset of features that can define all basic concepts for each object. Another words, a r-reduct is utilizing part of input features to uniquely identify output feature for each object. Furthermore, each r-reduct represents a decision rule.

Consider a single-feature r-reduct with four input feature and one output feature:2 x x x 1 , the entry 'x' mean that corresponded feature don't affect the determination of the feature output, only entry'2' does, and the decision rule can be expressed as:

If F1=2 Then d=1

For two-feature r-reduct:1 x 2 x 2 , only features F1 and F3 affect the determination of the output feature , and the decision rule is:

If F1=1 ? F3=2 Then d=2

Usually, there exist more than one r-reduct for each object. However, the knowledge rule of data not only can be inducted from lower and upper approximation of R.S.T but also can be directly analysed from all r-reducts.

III Modified rule generation algorithm for incomplete systems -MRGI

The MRG algorithm proposed in Guo. & Chankong can generate all possible rule-reducts in complete information systems including inconsistent data. However, it cannot deal with incomplete information systems. When dealing with IS, we have to transform the IS to S.

From other researches, we know that both methods (removing and replacing unknown information) have their drawbacks. Hence, it is necessary to develop an algorithm to induce rules directly from the original IS, which means: extract certain rules without changing the size of initial information systems.

Based on the MRG algorithm generating the minimal set of r-reducts, a generalized rule-generation algorithm for incomplete information systems (MRGI) is developed as follows:

There was 8 steps (Figure 1) by generalized MRG algorithm for incomplete information system.

Conclusions

In this article, a new rule-generation algorithm(MRGI) based on rough set theory has been proposed to generate the minimal set of rule-reduct which also represent the certain rules from incomplete information system. Some examples have been illustrated to demonstrate the validation of the methods we presented.

Compared with other complex rough sets based rule induction approachs, MRGI present not only the capacity to deal with uncertainty or inconsistent information but the high ability to generate concise and simple rules as well, and makes it easier to analyze and induce knowledge from large incomplete information systems. Without changing the size of the original system or adding possible values to the null values, the technique proposed in this paper generate knowledge rules directly from the original incomplete information systems.

We strongly believe that the rule induction technique presented in this paper will provide an efficient access for rule generation and knowledge induction, and will play a competent role to complement other existed rule induction Techniques for incomplete information systems.

Table 1. Medical diagnosis data (this table changed from [2])

Row no.	Attributes: Temperature	Dry-cough	Headache	Decision: Influenza
0	Missing	Absent	Absent	Absent
1	Normal	Absent	Present	Absent
2	Subfebrile	Absent	Present	Present
3	Subfebrile	Missing	Absent	Absent
4	Subfebrile	Present	Absent	Present
5	High	Absent	Missing	Absent
6	High	Present	Absent	Absent
7	High	Present	Absent	Present
8	High	Present	Present	Present
9	High	Present	Present	Present

Table 2. Decision table

Obj	F1	F2	F3	F4
x_0	M	0	0	0
x_1	0	0	1	0
x_2	1	0	1	1
x_3	1	M	0	0
x_4	1	1	0	0
x_5	2	0	M	0
x_6	2	1	0	0
x_7	2	1	0	1
x_8	2	1	1	1
x_9	2	1	1	1

Table 3. the results

U/A	a_1	a_2	a_3
$\{x_0\}$	M	0	0
$\{x_1\}$	0	0	1
$\{x_2\}$	1	0	1
$\{x_3\}$	1	M	0
$\{x_4\}$	1	1	0
$\{x_5\}$	2	0	0
$\{x_6, x_7\}$	2	1	0
$\{x_8, x_9\}$	2	1	1

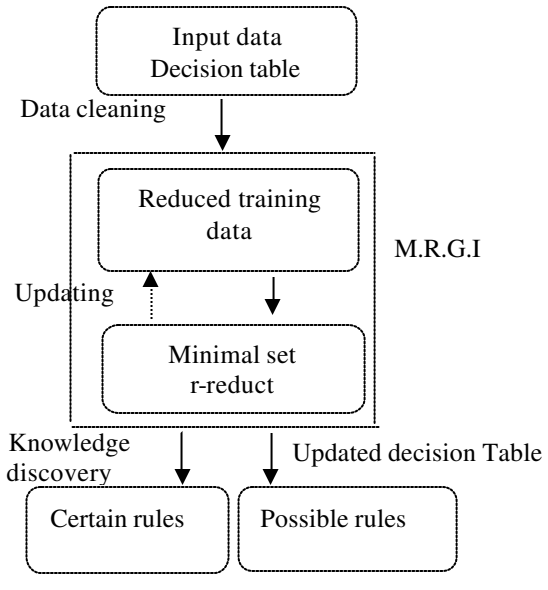

Fig.1 Mechanism of RGIPI including MRGI.