Data Mining Based on Rough Sets and Its Application in Risk Decision-making

LI WANQING, MA LIHUA
Hebei University of Engineering, Handan, 056038
CHINA
E-mail: malihua2004@126.com

WEI DONG
Hebei University of Engineering, Handan, 056038
CHINA
E-mail: wdongau@yahoo.com.cn

Abstract: The risk decision-making is an important aspect in the management practice. In the risk decision process of a project decision-making, it is necessary to use the algorithm to discover valuable knowledge and make a right decision. In the paper, a data mining method called Rough Sets is introduced in the field. And the algorithmic process of data mining based on Rough Set is studied. According to the Rough Set theory, firstly, the factors set is established including condition attribute and decision attribute. Secondly, experts qualitatively describe risk factors and establish a decision database, called decision table. Thirdly, the attribute reduction algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value of decision table. Fourthly, the minimum decision rules are abstracted based on data mining technology. Finally, the process of risk decision based on data mining of Rough Sets is analyzed in a case study.

Key-Words: Data mining, Rough Sets, minimum decision rule, attribute reduction, risk decision, project decision-making

1 Introduction

Data mining is a kind of method to process massive data and to find out some implied rules that are useful to make decisions [1]. Rough Sets theory proposed by Z. Pawlak in 1980s is one of such techniques. It is a novel mathematic method to study uncertain data, deficiency of data, incomplete data, or even inconsistent data [2]. And Rough set theory is very broad application area, such as expert systems, decision support systems, machine learning, pattern recognition, data mining, artificial intelligence and so on [3]. There is much uncertain information in the risk decision-making, such as project risk decision-making. Project decision-making is a high-risk project because of many uncertain causes, including complex technology, specialized equipment, special environment personnel disposition and so on [4]. How to control and decrease the risk is a difficult problem [5]. The process of decision-making is making right decision through right information and right way. So information plays an important role in the course of decision-making process. Information of proper quantities is necessary to make decision, that is to say, information of quantitative and qualitative influences directly the result of decision [6]. As far as the deciders are concerned, they hold the massive information in a project decision-making. Then it is necessary to use the algorithm to discover valuable knowledge and make right decision. Rough Sets theory applies to data mining supplying the mathematics tool for dealing with uncertain knowledge [7]. Liu Qing and Zeng Huanglin studied the characteristic and application of Rough Sets theory [8,9]. Yang Shanlin discussed the process of data analyzing based on data mining of Rough Sets, and proposed the application of this method to decision support system [10].

The rest of the paper is organized as follows. In section 2, data mining based on Rough Sets are introduced, including some concepts of Rough Sets theory, reduction algorithm based on Rough Sets, and the computational process. In section 3, a process of data mining is analyzed in a project decision-making. In this course, first, the set of factors is established, including condition factor and decision factor. Secondly, experts qualitatively describe risk factors and establish a decision knowledge database, called decision table. Thirdly, the algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value from the decision table. Fourthly, the minimum decision rules are created based on data
mining technology. Finally, the meaning of minimum decision rules is analyzed in a case study.

2 Data Mining Based on Rough Sets

2.1 Some concepts of Rough Sets theory

2.1.1 Indiscernibility relation
In Rough Sets, the relation is close between knowledge and classification, and knowledge is defined as an ability to classify. Suppose $K = (U, R)$ is a knowledge base, where $U$ is a nonempty finite set called domain, $R$ is the equivalence relations of $U$, $U/R$ is all the equivalence classes of $R$. $[X]_R$ is an equivalence class of $R$ including element $x \in U$. If $P \subseteq R$ and $P \neq \Phi$, then all intersection of equivalence relations are an equivalence relation in $P$, called indiscernibility relation about $P$, as in $\text{ind}(P), [x]_{\text{ind}(P)} = \bigcap_{R \in P} [x]_R, P \subseteq R$.

2.1.2 Upper approximation, lower approximation and boundary of Rough Sets
In Rough Sets, accuracy concepts are signified by two accuracy sets including upper approximation and lower approximation. In a knowledge based on $K = (U, R)$, for each subset $x \in U$ and an equivalence relation $R \in \text{ind}(K)$, suppose two subsets are as follows.

$R^+(X) = \{x \in [x]_R \subseteq X, x \in U\}$

$R^-(X) = \{x \in [x]_R \cap X \neq \phi, x \in U\}$

Then $R^+(X)$ and $R^-(X)$ are the upper and lower approximation sets of $X$ about $R$. Suppose boundary domain of $X$ about $R$ is $b_{R^+}(X) = R^-(X) - R^+(X)$. And suppose $\text{pos}R(X) = R^+(X)$ is the positive region of $X$ about $R$, $\text{neg}R(X) = U - R^+(X)$ is the negative region of $X$ about $R$.

2.1.3 Information system and decision table
In Rough Sets, the information system takes the form of relation table. Knowledge system with condition attribute and decision attribute is a decision table. A decision table is a kind of critical knowledge system. Suppose $S = (U, A, V, f)$ is a knowledge system, where $S = (x_1, x_2, \ldots, x_n)$ is a finite set of object, $A = (a_1, a_2, \ldots, a_n)$ is a finite set of attribute, here in $V$ is field composed of attribute $A, f : U \times A \rightarrow V$ is an information function, each element of $U$ with a unique value that is $a$ about $V$, $A = C \cup D$, $C$ is the condition attribute set, $D$ is the decision attribute set.

2.2 Reduction algorithm based on Rough Sets
Simplified table is the result of simplifying condition attribute, and the classification function remains to be. And simplified decision table contains less complicated condition attributes. We know a simplified condition is necessary in making decisions. The algorithm has 2 steps, i.e. attribute reduction and attribute value reduction as follows.

2.2.1 Attribute reduction
For an information system $S = (U, A, V, f)$, $A = C \cup D, B \subseteq C$, if $\gamma^c(D) = \gamma_b(D)$ and $B$ is individual in relation to $D$, then $B$ is the simplification of attribute $D$ in relation to $C$, as in $\text{RED}_D(C)$. The calculation is shown as follows.

Input: $C$, $D$, and $U$

Output: attribute reduction $C$ in relation to $D$

Step 1 $s \leftarrow 0, \text{RED}(s) \leftarrow \phi$;  

Step 2 $i \leftarrow 1$;  

Step 3 $j \leftarrow 1, m \leftarrow 0$;  

Step 4 For subset $C(i, j)$ of $C$, covering $j$ subset of element $i$

(1) $t \leftarrow 0$  

(2) If $(\text{RED}(t) \neq \phi) \land (\text{RED}(t) \subseteq C(i, j))$,

$m \leftarrow m + 1$, if $m = C[i]$, turn to Step 7, else turn to Step 5  

(3) If $t \geq s$ turn to (5)  

(4) $t \leftarrow t + 1$, turn to (2)  

(5) If $\gamma^c(D) = \gamma_{C(i, j)}(D)$ turn to (6), else turn to Step 5  

(6) $s \leftarrow s + 1$, $\text{RED}(s) \leftarrow C(i, j)$  

Step 5 If $j \geq C[i]$, turn to Step 6, else $j \leftarrow j + 1$, turn to Step 4  

Step 6 If $i \geq |C|$ ends, else $i \leftarrow i + 1$, turn to Step 3  

Step 7 Output $\text{RED}(s)$

2.2.2 Attribute value reduction
For in information system $S = (U, \text{RED}_p(C) \cup D, V, f)$, the calculation is shown as follows.
3 Case Study
The risk decision-making is a process from identification to settlement during the course of decision. And the process of data management contains 3 stages, such as collection, process and application.

3.1 Condition attribute sets and decision attribute sets
The risk factor sets are called condition attribute sets, which reflect the risks in project decision-making, including technical feasibility \( a \), amount of investment \( b \), capital-raising ability of project \( c \), and market expectation \( d \). Decision attribute sets include enterprise scale \( e \) and risk process methods \( f \). We can get information to make decision from a decision table, called risk decision table. The table is composed of rows and arrays to represent attributes and objects. We study a project decision-making practice to abstract Table 1 as follows.

3.2 Dispersing attribute and establishing knowledge base
Dispersing condition and decision attributes are used to establish the knowledge base. Firstly, using condition attribute sets from above; we disperse the results of expert evaluation as follows. Technical feasibility is divided into 3 grades \( \{1,2,3\} \) to represent \{low,average,high\}. Similarly, the amount of investment is also divided into 4 grades \( \{1,2,3,4\} \) to represent \{lower,low,high,higher\}. Capital-raising ability of project is divided into 2 grades \( \{1,2\} \) to represent \{bad,good\}. Market expectation is divided into 3 grades \( \{1,2,3\} \) to represent \{bad,average,good\}. Secondly, we use decision attribute sets above, and the results of expert evaluation are represented as follows. Enterprise scale is divided into 2 grades \( \{1,2\} \) to represent \{small,big\}. Risk process methods are divided into 3 kinds \( \{1,2,3\} \), including risk bearing, risk sharing and risk avoiding.

3.3 Attribute reduction
From Table 1, the redundant attributes are eliminated and core attributes are preserved. The minimum decision rules are composed of core attributes without redundant attributes. The new table is called reduced attribute table as follow.

3.4 Attribute value reduction
From Table 2, we get attribute value reduced table as follows.
3.5 Interpretation analysis

From Table 3, we know decision rules as follows.

\[ a_2d_2 \rightarrow (2,3) \]

or \[ a_2d_1 \rightarrow (2,3), a_1 \rightarrow (1,3), d_3 \rightarrow (2,2) \]

or \[ c_1 \rightarrow (2,2), c_2 \rightarrow (2,1) \]

From above, we know 4 decision rules as follows.

1. \[ a_2d_2 \lor a_2d_1 \rightarrow (2,3) \]
2. \[ a_1 \rightarrow (1,3) \]
3. \[ d_3 \lor c_1 \rightarrow (2,2) \]
4. \[ c_2 \rightarrow (2,1) \]

From above, we know 4 interpretative rules as follows.
1. If technical feasibility=average and market expectation=average (or bad) then big enterprise=risk avoiding.
2. If technical feasibility=low then small enterprise =risk avoiding.
3. If market expectation=good or capital-ability=bad then big enterprise=risk sharing.
4. If capital-raising ability=good then big enterprise =risk bearing.

From 4 decision rules above, we get strategy of risk decision as follows.
1. When the technical feasibility is “average”, and market expectation is “average” or “bad”, the strategy of big enterprise is risk avoiding. That is to say, that will give up the project or modify it.
2. When the technical feasibility is “low”, the strategy of small enterprise is risk avoiding. That is to say, that will give up project or modify it.
3. When market expectation is “good” or capital-raising ability of project is “bad”, the strategy of big enterprise is risk sharing. That is to say, that will share risks and profits with cooperators.
4. When capital-raising ability of project is “good”, the strategy of big enterprise is risk bearing. That is to say, that will solely bear the risk in full.

4 Conclusions