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

LI WANQING, MA LIHUA, WEI DONG
School of Economics and Management
Hebei University of Engineering
Guangming South Street 199,Handan,056038
CHINA
malihua2004@126.com,wdongau@yahoo.com.cn

Abstract: -In order to solve the problem of the redundant information to distinguish in the risk decision-making, in this paper, the data mining algorithms based on Rough Sets is studied. And we know 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 Sets 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.Plawlak 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 Sets 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. And There are many treatment of uncertain information, such as ANN,WNN and Wavelet analysis[4,5].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 [6]. How to control and decrease the risk is a difficulty problem [7]. 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 [8]. 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 [9,10]. Liu Qing and Zeng Huanglin studied the characteristic and application of Rough Sets theory [11,12]. 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 [13].

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, the computational process, an increm ental algorithm for attribute reduction and the reduction algorithm of deleting the rear attribute. 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 Data mining principle based on Rough Sets

2.1.1 Knowledge discovery in database (KDD) and data mining

Along with the data's increasing growth, some large-scale databases already have gone far beyond the degree which artificial could analyze, but KDD is an effective way to solve the problems above. KDD is a newly-involved research area which is based on the combination of artificial intelligence and machine learning technology. As a decision-making support process, KDD, this is mainly based on artificial intelligence, machine learning, pattern recognition, and statistics, highly-automated analysis massive data, data mining, and thus makes correct decisions. The most important step of the KDD processes is the data mining and the goal of data mining which is to mine the concealed and significant knowledge from the massive data .The data mining processes are shown as follows Fig.1.

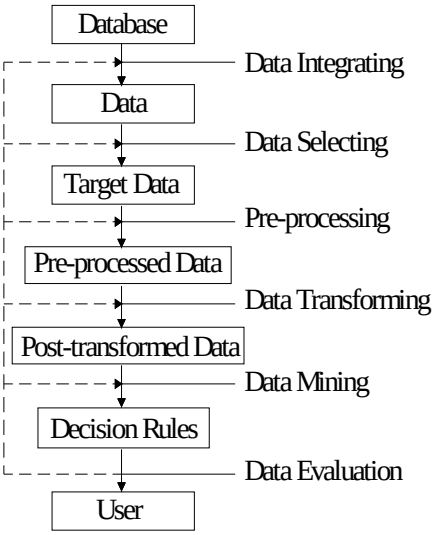

Fig.1. The procedure of data mining

From Fig.1, we know that the procedure of data mining contains 3 stages, data preparation, data mining, and data evaluation. And the stage of data preparation includes 4 steps, such as data integrating, data selecting, pre-processing and data transforming. The decision rules are abstracted by data mining. All the steps have connection with data evaluation.

2.1.2 The data mining process based on Rough Sets

The data mining process based on Rough Sets contains 5 stages, original data, discretization, knowledge base, algorithm of Rough Sets and minimum decision rule. The procedure is shown as follows Fig.2.

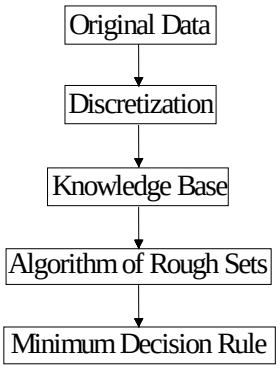

Fig.2. The process of data mining based Rough Sets

2.1.3 The data mining system

The data mining system is divided into 3 stages, including user interface, data mining, and source data. Man-machine interaction system is included in user interface. Data mining system is contained in the mining stage. Database system and knowledge base compose of the stages of source data. The structure is shown as following Fig.3.

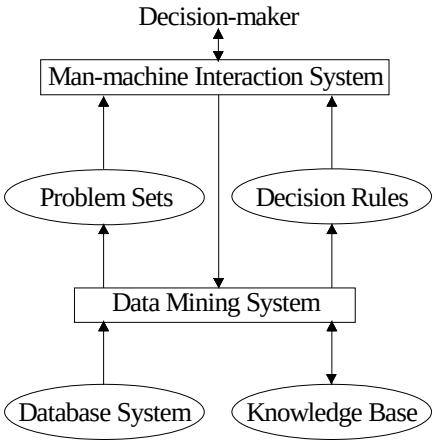

Fig.3. The structure of data mining system

2.2 Some concepts of Rough Sets theory

2.2.1 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 $ind(P)$, $[x]_{ind\{R\}} = \bigcap_{R \in P} [x]_R$, $P \subseteq R$.

2.2.2 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 $R \in ind(K)$, suppose two subsets are as follows.

$$R_{-}(X) = \{x | [x]_R \subset X, x \in U\}$$

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

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 $bn_R(X) = R^{-}(X) - R_{-}(X)$. And suppose $posR(X) = R_{-}(X)$ is the positive region of X about R, $negR(X) = U - R_{-}(X)$ is the negative region of X about R.

2.2.3 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 $S = (x_1, x_2, \cdots, x_n)$ is a finite set of object, $A = (a_1, a_2, \cdots, a_n)$ is a finite set of attribute, here in V is field composed of attribute A, $f : U \times A \to V$ is an information function, each element of U with a unique value that is a about V, $A = C \bigcup D$, C is the condition attribute set, D is the decision attribute set.

2.3 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.

2.3.1 Attribute reduction

For an information system $S = (U, A, V, f)$, $A = C \bigcup D$, $B \subseteq C$, 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 $RED_D(C)$. The calculation is shown as follows.

Input: C, D, and U

Output: attribute reduction C in relation to D

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

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 $(RED(t) \neq \phi) \wedge (RED(t) \subseteq C(i, j))$, $m \leftarrow m + 1$, if $m = C^i_{|C|}$, 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) $s \leftarrow s + 1$, $RED(s) \leftarrow C(i, j)$

Step 5 If $j \geq C^i_{|C|}$ turn to Step6, else $j \leftarrow j + 1$, turn to Step4

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

Step 7 Output $RED(s)$

2.3.2 Attribute value reduction

For in information system $S = (U, RED_D(C) \cup D, V, f)$, the calculation is shown as follows.

Input: $S = (U, RED_D(C) \cup D, V, f)$, $RED(C) = \{C_1, C_2, \cdots, C_n\}$

Output: core value table S' of S

Step1 $S' = (U, C \cup D, V' \leftarrow Null, f')$

Step2 For each condition attribute C_k (repeat as follows)

For each $x_i \in U$ and $C'_k(x_i) = Null$ (repeat as follows)

If

$$\exists x_i ((x_j \neq x_i) \wedge \forall C_l (C_l \neq C_k \wedge C_l(x_j) = C_l(x_i)) \wedge (D(x_j) \neq D(x_i))$$

Then $C'_k(x_j) = C_k(x_j)$, $C'_k(x_i) = C_k(x_j)$

Step3 Output S'

2.4 An increm ental algorithm for attribute reduction

2.4.1 The basic conception

(1) Information Systems

Information of the object studied is given by information system in the form of list, the rows of table corresponding to the object and the columns of table corresponding to the object's properties. Information systems can be expressed as a four-tuple $S=(U,A,V,f)$, in which U is the non-empty finite set of the object, called the domain; A is a non-empty finite set of property, $V=\underset{a\in A}{U}V_a$, V_a is the range of property a, $f:U\times A\to V$ is the information function, that is, $a\in A$, $x\in U$, $f(x,a)\in V_a$. The knowledge representation system with conditional attributes and decision attributes is called decision table.

(2) Relative and absolute distinction matrix

Inspired by the literature this paper gives the definition of the relative and absolute distinction matrix.

Definition 1. The definition of a relative distinction matrix is defined as a matrix B containing $mn(n-1)/2$ elements, where n is the number of the domain object and m is the number of condition attributes. The matrix elements are defined as follows:

$$B((i,j),k)=\begin{cases}1 & c_k(x_i)\neq c_k(x_j)\wedge d(x_i)\neq d(x_j)\\ 0 & other\end{cases}$$

$$1\le k\le m \quad i<n \quad j\le n \quad i<j$$

$B((i,j),k)=1$ represents that the first k attributes can distinguish the first i objects from the first j objects with the different decision-making value; $B((i,j),k)=0$ represents the first k attributes can not distinguish the first i objects from the first j objects. In fact matrix B is a $0-1$ matrix.

Definition 2. The absolute distinction matrix can be obtain by the relative distinction matrix and ,using B_1 said. Its element is defined as:

$$B((i,j),k)=\begin{cases}-1 & \forall 1\le p\le m, B((i,j),p)\\ & =0\wedge B_1((i,j),k)=d(x_i)\neq d(x_j)\\ 1 & B((i,j),k)=1\\ 0 & other\end{cases}$$

$$1\le k\le m \quad i<n \quad j\le n \quad i<j$$

Matrix B_1 and B are the same order, which contain the same number of elements.

(3)Condition attribute price P

Condition attribute price P refers to the sum of the cost, time, degree of difficulty and the size of risk required by getting a certain attribute values in real life. We say that $P_a\le P_b$ represents that the needed cost of obtaining property a is less than property b.

2.4.2 The basic idea

Usually we always hope that the higher property cost in the original attribute reduction is firstly replaced by the newly added property and the newly added property can distinguish between incompatible objects in the original system , so sorting the original attribute reduction with the price of attributes or the actual need decreasing, and then determine whether the newly added properties are the core attributes, and finally determine whether the newly added properties can replace attributes of the original reduction to get a new attribute reduction. The concrete steps is below, the algorithm is not only applicable to the situation of adding a new property, but also the situation of adding a number of new properties one time.

2.4.3 The specific steps

Input: $S=(U,A,V,f),C,N_i(i=1,2,\cdots m)$

Here C indicated an attribute reduction of the original system, the property in C has been sorted by property price; N_i is the value that each object is in the first i newly added property, expressed as a column vector, m is the total number of the newly added properties, S is a system after the original reduction.

Output: a meaningful attribute reduction in the new database, with the values of C and h represented.

Step1.Build relative distinction matrix B and absolute distinction matrix B_1 of system S according to the definition.

Step2.Initialize $i=1$.

Step3.According to the definition 1, we get a systematic relative discernibility matrix NB_i composed of the first i newly added property and the original decision attribute, actually as a vector, $(i=1,2,\cdots,m)$.

Step4.Identify the line that the elements is '−1' in B_1, denoted by $\{r_1,r_2,\cdots,r_k\}$,and remove the elements $(z_1,z_2,\cdots,z_k)$ that these lines $\{r_1,r_2,\cdots,r_k\}$ corresponding to from the NB_i, and if there is at least one element that is '1' in $\{z_1,z_2,\cdots,z_k\}$, this indicates that the first i-added elements is nuclear shown as $h_i=1$, or else $h_i=0$

Step5.Initialize $j=1$.

Step6.If $\left(NB_i-\left(2\times B(:,j)-sum(B')'\right)\right)\geq 0$, execute $C(j)=0;B(:,j)=0;$ otherwise $C(i,j)\neq 0$. Here, $2\times B(:,j)$ indicates that each element of j column in B multiplies by 2 ; $sum(B')'$ represent the sum of column B; $B(:,j)=0$ says changing the first j column in B into a '0' vector; $C(j)=0$ means that the first j attributes in C can be replaced by the added properties.

Step7.Supposing $j=j+1$, if $j\leq t$, go to Step6, otherwise go to Step8. t is the number of attributes in the original reduction C.

Step8.Output C and h . $h=1$ says the new database reduction is composed of the new attribute and the properties represented by the elements that is not '0' in C ; $h=0$ means if there appears '0' element in C , the new database reduction is composed of the new attribute and the properties represented by the elements that is not '0' in C ; when $h=0$, if not appear element '0' in C, then the original reduction is the new database reduction.

Step9.Supposing $i=i+1$, if $i\leq m$, go to Step4.

Step10.Supposing $Q=C-M=(Q_1,Q_2,\cdots,Q_u),t-(i-1)\leq u\leq t$, the elements in C and Q are in descending order according to property price, and let the set formed by attributes that are replaced by the new properties before the first $(i-1)$ in C the $M=(b_1,b_2,\cdots,b_s)$, then $s\leq i-1$, initializing $j=1$.

Step11.If $\left(NB_i-\left(2\times B_2(:,j)-sum\left(B_2{}'\right)'\right)\right)\geq 0$, execute B_2; $B(:,j)=0$; otherwise $Q(j)\neq 0$. B_2 is the new matrix obtained by the original relative distinguish matrix B getting rid of the '0' column vectors. Step12.Supposing $j=j+1$, if $j\leq u$, go to Step11, otherwise, go to Step8.

2.4.4 The verification and analysis of algorithms

Example 1 In information system $S=(U,A,V,f)$, set:

$U=\{1,2,3,4,5,6,7\}$,

$A=C\cup D$ $D=\{e\}$ $C=\{a,b,c,d\}$ $V_a=\cdots=V_e=\{0,1,2\}$.

An information system is shown in Table 1.

Table 1 The information system

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

We know after the reduction that $\{b,c,d\}$ is the attribute reduction of information systems, expressed as $C'=[1,2,3]$. Setting g_1,g_2 the two new properties, the values of each object in g_1,g_2 on are as follows:

$$N_1=[1100102]\quad N_2=[0002011]$$

Assume that $P_c\geq P_b\geq P_d$, that means $C=[2,1,3]$. According to the algorithm in the text: $h_i=1$, $C_1=[0,1,3]$; $h_2=0$, $Q_2=[1,0]$. Then $C=[0,1,0]$, so the additional attributes g_1 is a the nuclear of the new decision table and can be replaced by properties c, and although the new property g_2 is not nuclear of new decision-making table but it can replace the properties d , and thus get a attribute reduction $\{g_1,b.g_2\}$ of the new decision-making table . This reduction in practice is a more practical reduction, and if use the original algorithm to recalculate the new decision table, we can obtain the same reduction $\{g_1,b.g_2\}$, but it takes a long time.

Example 2 Using information systems that contains 238 objects, six attributes to verify, there are more advantages in its speed. Adding to a property, if re-use an algorithm such as ARBGS to reduce the new system, it takes time 4.828000 s, while the proposed algorithm above only needs 1. 75000 s time, and if relative distinction matrix and absolute distinction matrix of the original information systems have been derived during the reduction process ,it

only takes 0.109000 s to get the reduction actually needed, as many attribute reduction algorithms require to calculate discernibility matrix.

2.4.5 Result analysis

The incremental attribute reduction algorithm in this paper can apply to not only the situation of adding a property, but also the situation increasing the number of properties at the same time. A useful attribute reduction for dynamic database can be got by the algorithm, and the quality of classification is generally superior to the original, so this study is of practical significance. Although the algorithm is proposed for the incomplete information system, but modifying the definition of the relative distinction matrix and absolute distinction matrix a little can be extended to the incomplete information system.

2.5 The reduction algorithm of deleting the rear attributes

2.5.1 The Origin of the algorithm

Information system in rough sets theory can be shown by sets including four elements $S=(U,A,V,f)$. U is a nonempty set remarking all the record in database. A means all the attributes in the database. Suppose the information table is also the decision table, the attribute in A can be further divided into condition attribute C and decision attribute D, $A=C\cup D$. V is a set of attribute values. f is a function about attributes and records, where the value of $f(e,a)$ determines that record e is about the value of attribute a.

The equivalence.

Supposing an attribute set like $B\subseteq A$ contented $IND(B)$, $IND(B)=\{(x,y)\in U\times U/a(x)=a(y)\}$, for each subset $a\in B$ and equivalence relation is Indiscernibility Relation. $U/IND(B)$ is the set of all the equivalence on $IND(B)$. $B(x)$ is the Equivalence of object x.

Definition 3 Reduction. In an information system, $S=(U,A,V,f)$,for $B\subseteq A$,suppose $IND(B)=IND(B-(a))$ is set up, attribute a in sets A is dispensable, or it is in- dispensible.Suppose any relation a is indispensible,the record is independent, or it is reliable and dependent.

Further more defining, suppose Q,P is independent and $IND(Q)=IND(P)$, Q is a reduce for relation sets P.All the dispensable relation sets in Q is called the core of P ,which is showed as $core(P)$.It is not hard to know there are more than one reduction, and the $core(P)=\cap red(P)$,where $red(P)$ is the sets of all the reduction of P.

Definition4.Dependence on the attribute. In the formula of $r(B,D)=\dfrac{|POS_B(D)|}{|POS_C(D)|}$, POS_BD is divided as a positive region according to attribute sets B while POS_CD is according to the whole attribute sets.

Definition5.Attribute significance. $sig(a,B,D)=r(B,D)-r(B-\{a\},D)$

Suppose U is theory domain, R is the equivalence relations of U .That $P\in R$ is a sufficient and necessary condition of a reduction should satisfy the contents as follows:

(1)For $\forall a\in P$, P is independent.

(2)For $\forall a\in R-P$, there is $sig(P|a)=0$.

2.5.2 Algorithm and analysis

(1) Deleting the rear attribute algorithms

This algorithm is mainly deleting the less important attribute based on the thought of attribute significance degree from the condition attributes, thus a reduction is acquired. If the attributes are in the multiple significance, remove the combination of several major properties using this algorithm. The fewer the number of equivalence classes, then the less the property right on the contribution of the domain partition. The algorithm does not require the calculation of core, saving time and reducing space, so it has its advantages.

Input: Decision table $T=\langle U,C\cup D\rangle$, C is the condition attribute sets and D is the decision attribute sets.

Output:a reduction of this decision table.

Algorithm: $R=C//C$ is a contain attribute.

$N=Card(R)$

$For\ I=1\quad to\quad N$

$\forall c\in R$, calculate $sig(c,C,D)$and rank them in descending order.

$//sig$ is the attribute significance, whose calculating is based on the definition 5.

$\varepsilon=\min c_i$ (if the number of the lest attribute is more than one, then choose the biggest combination number as ε) .

$R=R-\varepsilon$.

If $r(R,D)=r(C,D)//R$ is the dependence of attributes,

Exit

Else

Loop

End if

End for

Return R

Then the sets R is a reduction of the decision table.

Comments: Suppose the number of attributes in decision table is m, the number of object is n. The significance of each attribute is needed to calculate in this algorithm and ranked in a certain sequence, eliminating the smallest degree each time.

So $O(mn^2+m\log m)$ is the time complexity. For mass data, it is relatively easy to calculate and avoids the waste of space and the bomb of the data combination.

(2)Algorithm analysis

The Algorithm makes use of the dependence of attribute $r(B,D)=\frac{|POS_B(D)|}{|POS_C(D)|}$,to judge whether to reduce. Since $|POS_C D|$ is a constant, whether the positive area $U|D$ will change is only considered when the object is classified after removing some attributes from the sets. $sig(a,B,D)=r(B,D)-r(B-\{a\},D)$ is used to calculating the significance of attribute for reducing. How to judge the importance of the attribute, whether the corresponding sort change after deleting it from the original attribute sets is only considered .If the classification changes, it says the attribute is important, else it is not important.

Proposition 1 R is a reduction according to the after deleting attribute algorithms. R is a reduction based on Definition1.

Sufficient conditions: If R is a reduction, for $\forall a\in R, POS_{R-\{a\}}(D)\neq POS_R(D)$ and R is independent.And the formula $r(R,D)=r(C,D)$ is simultaneously found, it is showed that the classification from R is the same as the equivalence class generated by C and $sig(R|a)=0$.

Necessary conditions: classification generated by R equal to the same equivalence class produced by $C \Rightarrow r(R,D)=r(C,D)$,then $sig(R|a)=0$. R is indepent.It shows that the classification will change when the attribute is deleted and can't be removed. So R is a reduction.

Two conditions are mutual restrained as well as satisfied each other simultaneously. R is a reduction according to the after deleting attribute algorithms. The testifies is over.

Proposition2. $O(mn^2+m\log m)$ is the time complexity.

Proof: Suppose the number of attributes in decision table is *m*, the number of object is n. In attribute sets, *m* divisions are calculated. The time complexity of each division is $O(|n^2|)$.So the object sets need be scanned twice in the worst condition. The object is scanned each time while the equivalence is calculated once. Thus the worst complexity is $O(|m| \ |n^2|)$. $O(m\log m)$ is the time complexity for the rapid ranking. Accordingly, $O(mn^2+m\log m)$ is the time complexity. The proof is completed.

(3)Examples

As is illustrated in the Table1 of Reference [14],Condition attributes $C=\{Outlook, Temperature, Windy, Humidity\}$,Decision attribute $D=\{Class\}$,the significance degree of the condition attributes based on category is found as follows:

$POS_{C-a}(D)=\{2,5,6,7,9,10,11,13\}$

$POS_{C-b}(D)=\{1,2,3,\cdots,14\}$

$POS_{C-c}(D)=\{1,2,3,\cdots,14\}$

$POS_{C-d}(D)=\{1,2,3,7,8,9,10,11,12,13\}$

$r(C-\{a\})=8/14, r(C-\{b\})=1, r(C-\{c\})=1,$

$r(C-\{d\})=10/14, r(C)=1$

They are calculated according to Definition 4, then:

$SGF(a,R,D)=6/14, SGF(b,R,D)=0,$

$SGF(c,R,D)=0, SGF(d,R,D)=4/14$

The significance of attribute Temperature and Humidity is the lest. Then the attribute Temperature which is the biggest composed number is picked as the ε based on the minimum number of attributes is bigger than one in the algorithms. Simultaneously, $\{R\}=\{a,c,d\}, r(R)=r(C)=1$, shows that the category of reduction sets doesn't change. Therefore, a reduction of the original data table is $\{Outlook, Humidity, Windy\}$.

(4) Result analysis

As the improvement and synthesize to "weighted moving average method"and "least squares method", they do preferable performance with possessing the advantages of easy-using and high computing rapid as well as

enhancing the accuracy. However, it is hard for this algorithm to avoid the impact of extreme data. Thus more complicate algorithms are needed in trend analyzing of complex timing sequence data, such as rising the time of interpolation and taking better interpolation method. Most of the data has a good smoothing effect, meanwhile, minimize the impact of the extreme data and noise and the algorithms has high efficiency. That is where the difficulty of reach is. If new type and excellent algorithms that satisfy the condition were put forward, trend analysis could make greater efforts in the field of judging decision where people play a bigger role.

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 The analysis of risk decision procedure based on data mining of Rough Sets

Risk control is a procedure from identification to settlement during the course of decision. And the procedure of data management contains 3 stages, collection, process and application. From Fig.1, a procedure of decision support system includes 5 sections, i.e. data collection, rough database, decision knowledge base, decision rules base, and decision interface. Rough Sets are essential in two steps, both of which are used from original database to knowledge base and from knowledge base to decision rules base.

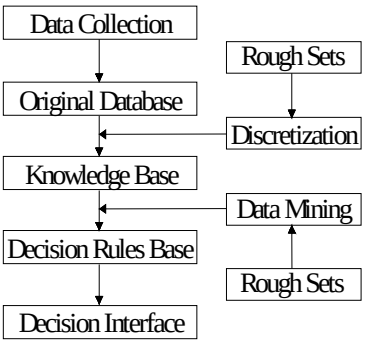

Fig.4. The analysis of risk decision procedure based on data mining of Rough Sets

3.2 Knowledge dependence and attribute weights

For the approximation space, $K=(U,R)$ and $P,Q\subseteq R$, if $ind(P)\subseteq ind(Q)$,then knowledge Q depends on the knowledge P, knowledge Q for P is defined as: $\frac{Card(Pos_P(Q))}{Card(U)}$,where $Card(\ )$ indicates that the set base, and $Pos_P[Q]$ is the positive region.

Suppose subsets C and the classification D from C ,if the category alters because of deleting the attribute a in sets C ,then the significant degree of a is as following :

$$Sig_{C-a}=\frac{r_C(C)-r_{C-a}(C)}{r_C(C)}=\frac{Card(Pos_C(D))-Card(Pos_{C-a}(D))}{Card(Pos_C(D))}$$

$$=1-\frac{Card(Pos_{C-a}(D))}{Card(Pos_C(D))}$$

The importance of each attribute is normalized to get the attribute weight W:

$$W_i=\frac{Sig_{C-a_i}(a_i)}{\sum_{j=1}^{n}Sig_{C-a_j}(a_j)}$$

3.3 Attribute reduction and attribute core

In the knowledge representation system, not all of the attributes are equally important, and even some attributes are completely redundant. Attribute reduction under the premise of keeping the ability of knowledge classification not changing deletes the redundant attributes, and finds the attribute core values.

The basic idea of attribute reduction is as follows: after getting rid of the attribute, if $Pos_C(D)=Pos_{C-a}(D)$, then it is claimed that attribute a is unnecessary to the decision attribute D, and its column can be removed, otherwise can't do it. Suppose all the attributes of a are not deleted, then C and D are indepent.It $B\subseteq C$ and B indepenes from D,then B is the reduction of C relative to D ,expressed as $\mathrm{Re}d_D(C)$ .All the intersection reductions of C relative to D are called the core, denoted as $Core_D(C)=\cap \mathrm{Re}d_D(C)$ .It should be noted that, the attribute reduction of C is not the sole, and core values may be empty.

3.4 Analysis of accessing to information and knowledge reasoning based on Rough Set

Knowledge reasoning process based on Rough Sets mainly includes the following steps:

(1)The establishment of related objects of a knowledge representation system.

(2)Attribute reduction for the knowledge representation system, and find the at- tribute core values.

(3)The reduction of attribute value.

(4)Rule-based reasoning and and simplify.

3.5 Condition attribute sets and decision attribute sets

The risk factor sets are called condition attribute sets, which reflect E-Commerce project risks, 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 an E-Commerce project to abstract Table 2 as follows.

Table 2. Risk decision table of E-Commerce project

U \ *A*	*a*	*b*	*c*	*d*	*e*	*f*
1	2	3	1	2	2	3
2	2	2	1	2	2	3
3	2	2	1	1	2	3
4	1	1	1	1	1	3
5	1	2	1	2	1	3
6	2	2	1	3	2	2
7	2	3	1	3	2	2
8	3	3	1	3	2	2
9	3	3	2	3	2	1
10	3	4	2	3	2	1
11	3	4	2	2	2	1
12	3	3	2	2	2	1
13	3	2	2	2	2	1

3.6 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.

3.7 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.

Table 3. Original table of Reduced attribute

U \ *A*	*a*	*c*	*d*	*e*	*f*
1,2	2	1	2	2	3
3	2	1	1	2	3
4	1	1	1	1	3
5	1	1	2	1	3
6,7	2	1	3	2	2
8	3	1	3	2	2
9,10	3	2	3	2	1
11,12,13	3	2	2	2	1

From Table 3, reduced attribute table is abstracted as follow.

Table 4. Reduced attribute table

U \ *A*	*a*	*c*	*d*	*e*	*f*
1	2	1	2	2	3
2	2	1	1	2	3
3	1	1	1	1	3
4	1	1	2	1	3
5	2	1	3	2	2
6	3	1	3	2	2
7	3	2	3	2	1
8	3	2	2	2	1

3.8 Attribute value reduction

From Table 4, we get attribute value reduced table as follows.

Table 5. Reduced attribute value table

U \ A	a	c	d	e	f
1	2	-	2	2	3
2	2	-	1	2	3
3	1	-	-	1	3
4	1	-	-	1	3
5	-	-	3	2	2
6	-	1	-	2	2
7	-	2	-	2	1
8	-	-	-	2	1

3.9 Interpretation analysis

From Table 4, we know decision rules as follows.

$a_2 d_2 \to (2,3)$

or $a_2 d_1 \to (2,3), a_1 \to (1,3), d_3 \leftarrow (2,2)$

or $c_1 \to (2,2), c_2 \to (2,1)$.

From above, we know 4 decision rules are abstracted as follows.

(1) $a_2 d_2 \vee a_2 d_1 \to (2,3)$.

(2) $a_1 \to (1,3)$.

(3) $d_3 \vee c_1 \to (2,2)$.

(4) $c_2 \to (2,1)$.

From above, we know 4 interpretative rules as follows.

(1)If the technical feasibility=average and the market expectation=average (or bad) then the big enterprise=risk avoiding.

(2)If the technical feasibility=low then the small enterprise =risk avoiding.

(3)If the market expectation=good or the capital-ability=bad then the big enterprise=risk sharing.

(4)If the capital-raising ability=good then the big enterprise=risk bearing.

3.10 Decision analysis

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

In this paper, some method of data mining based on Rough Sets on decision-making is studied. And we know that the project risk decision is sort of multiple attribute decision-making processes. It is certain that we have to deal with massive data. So data mining technology is necessary, for it can abstract implicit regularity from massive data. In this paper, the reduction algorithm based on Rough Sets is proposed as a practical data mining technology. A new decision method is proposed in order to solve the risk decision problem of a project. The factor set is established including condition attribute and decision attribute. Then experts qualitatively describe risk factors and create a decision table, and the attribute reduction algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value of decision table. Finally, the minimum decision rules are created based on data mining results. And the rules with explain meanings are abstracted in the case study. We can make proper decision from the rules to improve precision and explanatory ability in project management practice.

References:

[1]Lv Anmin, Lin Zongjian, Li Chenming. Approaching of Data Mining and Knowledge Discovery [J]. *Survey Science*, Vol.12, No.4, 2000, pp. 36-39. (in Chinese)

[2]Z. Pawlak. Rough Sets.*International Journal of Computer and Information Sceinces*, 1982, (11): 341-356.

[3]Vasant Dhar, Roger Stein. Intelligent Dicision Support Methods. *The Science of Knowledge Work*. Printice Hall, 1997.

[4] Wu Zhongli,Wei Liao,Pu Han. Application of Wavelet Network for Detection and Localization of Power Quality Disturbances[J].*WAEAS Transactions on Power Systems*, Vol.1, No.10, 2006, pp. 2029-2034.

[5] Wu Zhongli,Wei Liao,Pu Han. Early Detection for Short-circuit Fault in Distributed Power System Based on Wavelet Fractal Exponent

Analysis[J].*WAEAS Transactions on Power Systems*, Vol.1, No.12, 2006, pp. 2035-2040.
[6]Lu Youjie,Lu Jiayi. *Project Risk Management* [M].Beijing: Tsinghua University Press, 1998. (in Chinese)
[7]Zhuang Enyue,Wang Mingzhu. *Risk Investment & Case Analysis*[M].Beijing: China Audit Press, 1999. (in Chinese)
[8]ZhouSanduo,ChenChuanming.*Management*[M]. Higher Education Press,2000. (in Chinese)
[9]Wu Bing, Li Weixiang, Zhao Lindu. Least Decision-Making Method Based on Rough Set Theory [J]. *Information Technique*, Vol.4, No.4, 2002, pp. 4-5. (in Chinese)
[10]Li Wanqing,Ma Lihua,Meng Wenqing. Analysis of Procedure of Project Decision Based on Data Mining of Rough Sets[J]. *WSEAS Transactions on Business and Economics*, Vol.3, No.10, 2006, pp. 661-666.
[11]Liu Qing. *Rough Sets and Rough Inference* [M]. Beijing: Science Press, 2001. (in Chinese)
[12]Zeng Huanglin. *Rough Sets and Application* [M].Chongqing University Press,1996. (in Chinese)
[13]Yang Shanlin. *Intelligent Decision Method and Intelligent Decision Support System* [M]. Science Press, 2005. (in Chinese)
[14]Chen Chuanbo,Pan Fei,Li Qishen etal. Trend Weighting Smooth Predicting Model in Time Serial[J]. *Mini-micro Systems*, Vol.22, No.11, 2001, pp. 1299-1301. (in Chinese)