A Novel Extension Data Mining Approach based on Rough Sets Pair Analysis

Zhi-hang Tang [1, 2]
[1] School of Computer and Communication, Hunan Institute of Engineering, Xiangtan, 411104, China
[2] Glorious-Sun School of Business and Management, Donghua University Shanghai, 200051, China
Email:tang106261@mail.dhu.edu.cn

Abstract— **In the data base of information system, usually there are some attributes which are unimportant to the decision attribute, and some records that disturb the decision making. In this paper, reducing the condition attributes based on the matter-element theory and rough set method, calculating the importance to the decision attribute for each condition attribute after reduction, and data mining the relevant rules based on the reduced attributes, extension relevant function is used to depict quality of data gather in data mining. Finally, how to tap new customers and how to recommend an appropriate brand to new customers, Research result indicates that extension data mining can provide effective decide support for the Decision-making of enterprise.**

Index Term— **rough set, extension data mining, attributes reduction, matter-element**

I. INTRODUCTION

Extenics was first brought out by the famous researcher Cai Wen in 1983 in China, the major goal of this subject is to solve the incompatible problems through studying the extension probability of things. Matter-element is the logic cell of extenics and puts the matter, the characteristics and their measure together into consideration [1-13].

Data mining [14, 15] has gained popularity in the database field recently; it has been mostly used by statisticians, data analysts and so on. Data mining techniques can be divided into five classes of methods: predictive modeling; clustering; data summarization and change and deviation detection [16, 17]. Some of these techniques are beginning to be scaled to operate on databases.

In this paper, reducing the condition attributes based on the matter-element theory and rough set method, calculating the importance to the decision attribute for each condition attribute after reduction, and data mining the relevant rules based on the reduced attributes, extension relevant function is used to depict quality of data gather in data mining. Extension data mining has advantage in the decision-making of the enterprise. Extension relevant function can be used to quantitatively depict data gather of data mining. It can effectively solve the problem of the data quality, and provide clear decision-making for the enterprise. It also can provide a new thinking mode and approach for data mining which extenics is applied in.

II. BASIC CONCEPTION

A. Rough sets

A.1. Introduction of rough sets theory

Rough Sets Theory, a mathematical theory for data analysis, was first introduced by Z．Pawlak in 1982. By defining knowledge from a new viewpoint, it can be used to solve uncertain and imprecise problems. The most special characteristic of this theory is that it doesn’t need any earlier or additional information to tackle questions besides some necessary data muster．In combination with neural network，expert system，fuzzy theory, evidence theory，or genetic algorithm, Rough Sets Theory is widely used in various fields[18], such as knowledge acquirement，data mining，pattern recognition，machine learning，and decision support．As an important component of Rough Sets Theory, Reduction of Attributes attracts increasing attention in both theory and application. Reduction of Attributes corresponds to problems on selecting subsets of attributes in machine learning. It can effectively reduce information redundancies，and help people make a correct and concise decision. In Reduction of Attributes of Rough Sets，the reduction of information system is usually not unique. The number of attributes directly affects the length of the coding of decision rules． To acquire the most concise decision rules，a reduction including the least attribute is required．However, Wong and Ziarko[19] have proved that the minimal reduction in an information system is a hard problem，so recently there is no efficient algorithm in finding an optimal reduction[20].

A.2. Approximations

X is a muster，and R is an equivalence relation．The lower approximations R of X is denoted by $R_*(X)$ and defined as below：

$$R_*(X) = \{X \subseteq U : R(X) \subseteq X\} \qquad (1)$$

The upper approximations R of X is denoted by $R*(X)$ with the definition:

$$R^*(X) = \{X \subseteq U : R(X) \cap X \neq \phi\} \qquad (2)$$

Rough Sets are defined by

$$BN_R(X) = R^*(X) - R_*(X) \qquad (3)$$

$$POS_R(X) = R_*(X) \qquad (4)$$

$$NEG_R(X) = U - R^*(X) \qquad (5)$$

A.3. Information system

Given an information system, $T = (U, C, D, V_a)$, where U is the discussion universe，C is a set of

doi:10.4304/jsw.5.4.447-454

condition attributes, D is a set of decision attributes, and V_a is a set of values of attributes.

In the crisp set, an element either belongs to or does not belong to a set, so the range of the truth-values is [0, 1], which can be used to solve a two-valued problem.

A.4. Dependence of attributes

The purpose to discuss Dependence of Attributes is to analyze the inner relation among studied data. In the theory about Rough Sets, the degree of dependence among attributes is measured by $\alpha_R(X)$, defined as bellow.

$$\eta_* = k(R_*(X))/k(U) \qquad (6)$$

$$\eta^* = k(R^*(X))/k(U) \qquad (7)$$

$$\alpha_R(X) = k(R_*(X))/k(R^*(X)) \qquad (8)$$

B. Fuzzy logic

Fuzzy Logic has rapidly developed into one of the most successful modern technologies in dealing with sophisticated control systems. It resembles human decision-making with an ability to generate precise solutions from either certain or approximate information. Fuzzy Logic also bridges an important gap in engineering design methods left vacant by purely mathematical approaches (e.g. linear control design), and purely logic-based approaches (e.g. expert systems) in the system design. In addition, the other approaches often require accurate equations to model real-world behaviors, while fuzzy design can accommodate the ambiguities of the real-world human language and logic. It provides an intuitive method for a systemic description in human terms and automates the conversion of those systemic specifications into effective models.

Fuzzification (using membership functions to graphically describe a situation). Selection of fuzzy language variables and membership functions by $\frac{f_{j1}^i}{R_{j1}} + \frac{f_{j2}^i}{R_{j2}} + \cdots + \frac{f_{jl}^i}{R_{jl}}$, produces fuzzification information.

Defuzzification (obtaining the crisp or actual results). Following the equation $f_{jk}^i = \max(f_{j1}^i, f_{j2}^i, \cdots f_{jl}^i) \quad 1 \le k \le l$, we can get defuzzification information.

C. sets pair analysis (SPA)

Sets pair [11] is a pair of two relative sets. SPA is a method to process various uncertainties according to the connecting degree $u = a + bi + cj$. The two relative sets may have three relationships, i.e., identical, different, and contrary, and the connecting degree is an integrated description of these relationships.

H = (*A* , *B*) is assumed to be a set pair of two sets *A* and *B*. For certain application, H has totally *N* attributes. *S* of them is defined as the mutual part of *A* and *B*, *P* of them are the contrary part, and the residual attributes *F* = *N* - *S* – *P*. Then the connecting degree of *H*

$$u(\omega) = \frac{S}{N} + \frac{F}{N} i + \frac{P}{N} j = a + bi + cj \qquad (9)$$

Where $\frac{S}{N}$ is the identical degree, $\frac{F}{N}$ is the different degree, and $\frac{P}{N}$ is the contrary degree.

D. Extension set

In contrast to the crisp set, the fuzzy set allows for the description of concepts in which the boundary is not explicit. It concerns not only whether an element belongs to the set but also to what degree it belongs. The range of membership function is [0, 1] in fuzzy set. The extension set extends the fuzzy set from [0, 1] to [-∞, +∞]. This means that an element belongs to any extension set to a different degree. Define the membership function by $k(x)$ to represent the degree an element belongs to a set. A degree between zero and one corresponds to the normal fuzzy set. When $k(x) < 0$, it describes the degree to x does not belong to a set, which is not defined in a fuzzy set. When $-1 < k(x) < 0$, this means that the element x still has a better chance to be included into the set if the set is adjusted. $k(x) < -1$ implies that the element x has no chance to belong to the set, it is also to represent the degree of an element not belonging to a set. The extension theory tries to solve incompatibility or contradiction problems by the transformation of the matter element. Comparisons of crisp sets, fuzzy sets, rough sets and extension sets are shown in Table 1.

(1) Definition of matter-element

Defining the name of a matter by O, one of the characteristics of the matter by c and the value of c by v, a matter-element in extension theory can be described as follows:

Table 1. Four different sorts of mathematical sets.

Compared item	Standard set	Fuzzy set	Rough set	Extension set
Research objects	Data variables	Linguistic variables	Incomplete information	Contradictory problems
Model	Mathematics model	Fuzzy mathematics model	Attribute Reduction	Matter-element model
Descriptive function	Transfer function	Membership function	Upper/lower approximation set	Correlation function
Descriptive property	Precision	Ambiguity	Imprecise	Extension
Range of set	$C_A(x) \in [0,1]$	$\mu_A(x) \in [0,1]$		$K_A(x) \in [-\infty,+\infty]$

$$M=(O,c,v) \tag{10}$$

Where O, c and v are called the three fundamental elements of the matter-element. For example, M = (Tang, Weight, 60 kg) can be used to state that Tang's weight is 60 kg. If the value of the characteristic has a classical domain or a range, we define the matter-element for the classical domain as follows:

$$M=(O,c,v)=(o,c,\langle v^l,v^u\rangle) \tag{11}$$

Where v^l and v^u are the lower bound and upper bound of a classical domain.

(2) Multi-dimensional matter-element

Assuming $M=(O,c,v)$ a multi-dimensional matter-element, $c=[c_1,c_2...,c_n]$ a characteristic vector and $v=[v_1,v_2...,v_n]$ a value vector of c, then a multi-dimensional matter-element is defined as where

$$M=(O,c,v)=\begin{bmatrix} O & c_1 & v_1 \\ & c_2 & v_2 \\ & \vdots & \vdots \\ & c_n & v_n \end{bmatrix}=\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} \tag{12}$$

Where $M=(O,c_i,v_i)\, i=(1,2,\cdots n)$ is defined as the sub-matter-element of M .

(3) Divergence of matter-element

A matter may have many characteristics; the same characteristics and values may also belong to some other matter. In extension theory, there are some formulations to express these points as follows:

Theorem 1. If a matter has many characteristics, it can be written as

$$M\prec(O,c,v)\prec\{(O,c_1,v_1),(O,c_2,v_2),\cdots,(O,c_n,v_n)\} \tag{14}$$

The symbol $"\prec"$ indicates the mean of the extension.

Theorem 2. If some other matter has the same characteristic, it can
be written as

$$M\prec(O,c,v)\prec\{(O_1,c,v_1),(O_2,c,v_2),\cdots,(O_n,c,v_n)\} \tag{15}$$

Theorem 3. If some matter has the same value, it can be written as

$$M\prec(O,c,v)\prec\{(O_1,c_1,v),(O_2,c_2,v),\cdots,(O_n,c_n,v)\} \tag{16}$$

(4) Definition of extension set

Let U be the universe of discourse, x is a discretional element in U , is a mapping from U to real number field R , $T=(T_U,T_k,T_x)$ is a given extension transformation.

Then

$$\tilde{E}(T)=\{(x,y,y')\,|\,x\in T_U U, y=k(x)\in R, y'=T_k k(T_x x)\in R\} \tag{17}$$

is called an extensible set in U , $y=k(x)$ is a dependent function of $\tilde{E}(T)$, $y'=T_k k(T_x x)$ is an extension function of $\tilde{E}(T)$. Among which T_U is the transformation of U , T_k is the transformation of the dependent function, T_x is the transformation of element x . The $k(x)$ maps each element of U to a membership grade between $-\infty$ and ∞ .The higher the degree, the more the element belongs to the set. Under a special condition, when $0<k(x)<1$, it corresponds to a normal fuzzy set. $k(x)<-1$ implies that the element x has no chance to belong to the set. When $-1<k(x)<0$, it is called an extension domain, which means that the element x still has a chance to become part of the set.

(5) Definition of distance

If $X_0=\langle a,b\rangle$ and $X \;=\langle c,d\rangle$ are two intervals in the real number field, and $X_0\subset X$, then the correlation function in the extension theory can be defined as follows:

$$k(x)=\begin{cases} \dfrac{\rho(x,X_0)}{D(x,X_0,X)}-1, & \rho(x,X_0)=\rho(x,X\;) \\ & and\ x\notin X_0 \\ \dfrac{\rho(x,X_0)}{D(x,X_0,X)} & others \end{cases} \tag{18}$$

Where

$$\rho(x,X_0)=|x-\frac{a+b}{2}|-\frac{b-a}{2} \tag{19}$$

$$\rho(x,X\;)=|x-\frac{c+d}{2}|-\frac{d-c}{2} \tag{20}$$

$$D(x,X_0,X\;)=\begin{cases} \rho(x,X\;)-\rho(x,X_0) & \rho(x,X\;)\neq\rho(x,X_0),x\notin X_0 \\ \rho(x,X\;)-\rho(x,X_0)+a-b & \rho(x,X\;)\neq\rho(x,X_0),x\in X_0 \\ a-b & \rho(x,X\;)=\rho(x,X_0) \end{cases} \tag{21}$$

The correlation function can be used to calculate the membership grade between x and X as shown in Fig. 1 When $k(x)>0$, it indicates the degrees to which x belongs to X_0 .

When $k(x) < 0$ it describes the degree to which x does not belong to X_0, which is not defined in fuzzy set theory. When $-1 < k(x) < 0$, it is called the extension domain, which means that the element x still has a chance to become part of the set if conditions change.

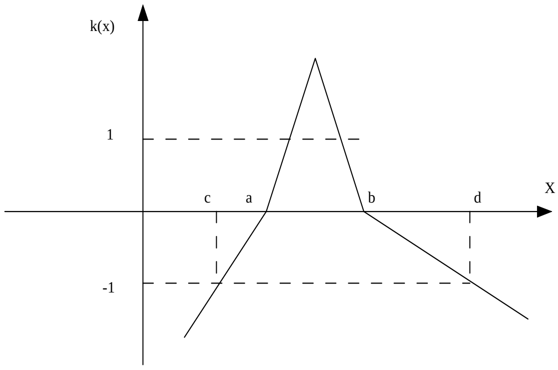

Fig. 1. The extended membership function

An extension set $\tilde{E}$ in U can be denoted by $\tilde{E} = E^+ \cup E_0 \cup E^-$

Where

$$E^+ = \{u | u \in U, k(u) > 0\} \quad (22)$$

$$E^- = \{u | u \in U, k(u) < 0\} \quad (23)$$

$$E_0 = \{u | u \in U, k(u) = 0\} \quad (24)$$

E^+ is called a positive domain in $\tilde{E}$, it can describe the degrees to which x belongs to X_0. E^- is called a negative domain in $\tilde{E}$, it describes the degree to which x does not belong to X_0. E_0 is called a zero boundary.

If $T \neq e$, then

$$\dot{V}_+(T) = \{u | u \in U, y = k(u) \le 0, y' = k(T_u u) > 0\} \quad (25)$$

is called the positive extension field of U as to T_u;

$$\dot{V}_-(T) = \{u | u \in U, y = k(u) \ge 0, y' = k(T_u u) < 0\} \quad (26)$$

is called the negative extension field of U as to T_u;

$$V_+(T) = \{u | u \in U, y = k(u) > 0, y' = k(T_u u) > 0\} \quad (27)$$

is called the positive stable field of U as to T_u;

$$V_-(T) = \{u | u \in U, y = k(u) < 0, y' = k(T_u u) < 0\} \quad (28)$$

is called the negative stable field of U as to T_u;

$$V_0(T) = \{u | u \in U, y' = k(T_u u) = 0\} \quad (29)$$

is called the extension boundary of U as to T_u

The partition of domain is shown in Fig2 and Fig3.

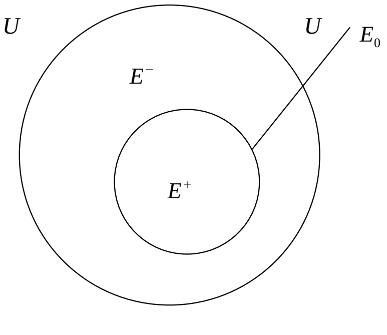

Fig 2.The partition of domain U when T=e

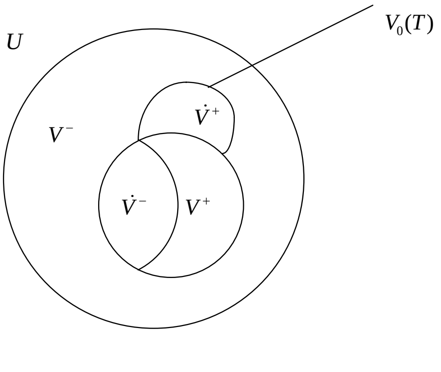

Fig 3. The partition of domain U when extension transformation

III. Extension Data Mining base on Rough Set

A. Attribute reduction based on rough set

We often face a question whether we can remove some data from a data table preserving its basic properties that is whether a table contains some superfluous data.

For example, it is easily seen that if we drop in Table 1 either the attribute Headache or Muscle-pain we get the data set which is equivalent to the original one, in regard to approximations and dependencies. That is we get in this case the same accuracy of approximation and degree of dependencies as in the original table, however using smaller set of attributes.

In order to express the above idea more precisely we need some auxiliary notions. Let B be a subset of A and let a belong to B.

- We say that a is dispensable in B if I(B) = I(B – {a}); otherwise a is indispensable in B.
- Set B is independent if all its attributes are indispensable.
- Subset B' of B is a reduct of B if B' is independent and I(B') = I(B).

Thus a reduct is a set of attributes that preserves partition. It means that a reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes. In other words, attributes that do not belong to a reduct are superfluous with regard to classification of elements of the universe.

Reducts have several important properties. In what follows we will present two of them.

First, we define a notion of a core of attributes.

Let B be a subset of A. The core of B is the set off all indispensable attributes of B.

The following is an important property, connecting the notion of the core and reducts

$$Core(B) = \bigcap Red(B) \qquad (30)$$

where Red(B) is the set off all reducts of B.

Because the core is the intersection of all reducts, it is included in every reduct, i.e., each element of the core belongs to some reduct. Thus, in a sense, the core is the most important subset of attributes, for none of its elements can be removed without affecting the classification power of attributes.

To further simplification of an information table we can eliminate some values of attribute from the table in such a way that we are still able to discern objects in the table as the original one. To this end we can apply similar procedure as to eliminate superfluous attributes, which is defined next.

- We will say that the value of attribute a∈B, is dispensable for x, if $[x]_{I(B)} = [x]_{I(B - \{a\})}$; otherwise the value of attribute a is indispensable for x.
- If for every attribute a∈B the value of a is indispensable for x, then B will be called orthogonal for x.
- Subset B' ⊆ B is a value reduct of B for x, iff B' is orthogonal for x and $[x]_{I(B)} = [x]_{I(B')}$.

The set of all indispensable values of attributes in B for x will be called the value core of B for x, and will be denoted $CORE^x(B)$.

Also in this case we have

$$CORE^x(B) = \bigcap Red^x(B) \qquad (31)$$

where $Red^x(B)$ is the family of all reducts of B for x.

Suppose we are given a dependency C ⇒D. It may happen that the set D depends not on the whole set C but on its subset C' and therefore we might be interested to find this subset. In order to solve this problem we need the notion of a relative reduct, which will be defined and discussed next.

Let C,D ⊆ A. Obviously if C' ⊆ C is a D-reduct of C, then C' is a minimal subset of C such that

$$\gamma(C, D) = \gamma(C', D). \qquad (32)$$

- We will say that attribute a∈C is D-dispensable in C, if $POS_C(D) = POS_{(C-\{a\})}(D)$; otherwise the attribute a is D-indispensable in C.
- If all attributes a∈C are C-indispensable in C, then C will be called D-independent.
- Subset C' ⊆ C is a D-reduct of C, iff C' is D-independent and $POS_C(D) = POS_{C'}(D)$.

The set of all D-indispensable attributes in C will be called D-core of C, and will be denoted by $CORE_D(C)$. In this case we have also the property

$$CORE_D(C) = \bigcap Red_D(C) \qquad (33)$$

where $Red_D(C)$ is the family of all D-reducts of C.

If D = C we will get the previous definitions.

$$IND(R) = IND(R - \{r\}) \qquad (34)$$

Where ind ( ) denotes the indiscernibility relation, and r R, which is the attribute sets. Obviously, if Eq. (34) holds, r is the redundant attribute element to describe the knowledge base characterized by attribute sets R. As a result, r can be removed from R, which is so-called knowledge simplification related to the classification problem. Moreover, the simplified attribute sets ind (R) is equivalent to the original attribute sets R, so some attributes can be reduced from the original Table.

B. Extension relevant rule

Relevant rule is defined that certain cases can bring about others cases. Such as rule $X \Rightarrow Y$, X and Y are the attribute variables in database. Extension relevant rule with matter-element is $\overset{n}{\underset{i=1}{\wedge}} r_i \Rightarrow (l)R$ .Relevant rules with combined type are rules which have essence-element item and extension transform item. It is $r_1 \wedge r_2 \wedge \cdots \wedge r_n \Rightarrow (l)R$. Relevant rule with combined type is fit for researching relevant rule of complicated system.

C. Decide classical field and modulation field

According to every characteristic variable, its data range can be acquired. Consequently classical field and modulation field of different levels, which is correlative with each characteristic, will be ensured.

$$M_{cf} = (O_{cf}, c, v) = \begin{bmatrix} O_{cf} & c_1 & \langle v_{cf1}^l, v_{cf1}^h \rangle \\ & c_2 & \langle v_{cf2}^l, v_{cf2}^h \rangle \\ & \vdots & \vdots \\ & c_n & \langle v_{cfn}^l, v_{cfn}^h \rangle \end{bmatrix} \qquad (35)$$

In formula, O_{cf} expresses the different level. $c_i (i = 1,2,\cdots,n)$ expresses the characteristic of O_{cf} . v_{cf} is the variable range which is ensured by characteristic variables $c_i(i=1,2,\cdots,n)$ of O_{cf} .So v_{cf} is called $\langle v^l_{cfi}, v^h_{cfi}\rangle(i=1,2,\cdots,n)$ which is a classical field. This is similar to $X_0 = \langle a,b\rangle$.

$$M_{mf} = (O_{mf}, c, v) = \begin{bmatrix} O_{mf} & c_1 & \langle v^l_{mf1}, v^h_{mf1}\rangle \\ & c_2 & \langle v^l_{mf2}, v^h_{mf2}\rangle \\ & \vdots & \vdots \\ & c_n & \langle v^l_{mfn}, v^h_{mfn}\rangle \end{bmatrix} \quad (36)$$

v_{mf} is the variable range which is ensured by characteristic variables of O_{mf} .So v_{mf} is called $\langle v^l_{mfi}, v^h_{mfi}\rangle(i=1,2,\cdots,n)$ which is a modulation field. This is similar to $X \quad = \langle c,d\rangle$.

D. Compute relevant degree according to relevant function

Let $I_i(i=1,2,\cdots m)$ be the subsets of the extension set O , $I_i \subset O, (i=1,2,\cdots,m)$ To any testing object $p \in P$, using the following steps to determine whether p belongs to the certain subset I_i ,and calculates the dependent degree.

Where $c_i(i=1,2,\cdots,n)$ are n different characteristics of I_i ,and v_{cf} are the range of $c_i(i=1,2,\cdots,n)$ associated with subset $I_i(i=1,2,\cdots m)$

Based on the analysis of characteristic variables, relevant function can be expressed as follows:

Relevant degree of the identified object O about the $j(j=1,2,\cdots,m)$ level is:

$$k_{ij} = \sum_{i=1}^{n} \alpha_{ij} \cdot \frac{k_j(x_i)}{\max|k_j(x_i)|}, \quad i=1,2,\cdots.n; j=1,2,\cdots \cdot m \quad (37)$$

$$k_i = \overset{m}{\underset{j=1}{\vee}} k_{ij} \quad (38)$$

α is Right weighted value. Determine the weighted value of each characteristic and calculate the value of dependent function. Here we introduce the proportion of the weighted value of each characteristic, calculated as:

$$\alpha_{ij} = \frac{x_j / b_{ij}}{\sum_{j=1}^{m} x_j / b_{ij}} \quad (39)$$

Finally, we determine the category of the testing sample. If $k_i = \max k_{ij} \ j=1,2,\cdots,m$,It means the testing sample belongs to I_i .

If $k_i \le 0 \ j=1,2,\cdots,m$ is right for any j , it means the testing sample is not belonging to any category that you have divided.

IV. APPLICATION EXAMPLE

China Unicom's current brand has Shijie Feng, Xin Shili and Ruyi Tong, how to tap new customers and how to recommend an appropriate brand to new customers? Their choices are Shijie Feng, Xin Shili, Ruyi Tong, or other?

A. Acquire and deal with data

There are four factors which affect the choices of the new customers. They contain age, basic call charge, message charge and GRPS charge. Above value of four factors can be acquired from survey of related groups. The paper analyses the data of the brand survey which is from more than 2000 customers, such as Table 2.

Table2. Data collection

customer	basic call charge(A)	message charge(B)	GRPS charge(C)	Phone model(D)	the brand(E)
1	100	20	50	Nokia	Xin Shili
2	340	2	5	Motorola	Shijie Feng
3	220	18	30	Nokia	Xin Shili
4	1000	5	14	Nokia	Shijie Feng
5	50	20	30	Lenovo	Xin Shili
6	100	5	14	Nokia	Ruyi Tong
7	460	2	5	Motorola	Ruyi Tong
…	…	…	…	…	…
amount to 2068 data					

We get Table 3 through $\frac{f^i_{j1}}{R_{j1}} + \frac{f^i_{j2}}{R_{j2}} + \cdots + \frac{f^i_{jl}}{R_{jl}}$

Table 3 customer information after Fuzzification

customer	basic call charge(A)	message charge(B)	GRPS charge(C)	Phone model(D)	the brand(E)
1	1/2	1/1	1/1	3	1
2	0.4/2+0.6/3	0.6/4+0.4/3	1/3	1	3
3	0.7/2+0.3/3	0.2/2+0.8/1	0.6/2+0.4/1	3	1
4	1/4	1/3	0.6/2+0.4/3	3	3
5	1/1	1/1	0.6/2+0.4/1	2	1
6	1/2	1/3	0.6/2+0.4/3	3	2
7	0.1/2+0.9/3	0.4/3+06/4	1/3	1	2
…	…	…	..	…	..

According to Table 3, we define $f^{i}_{jk} = \max(f^{i}_{j1}, f^{i}_{j2}, \cdots f^{i}_{jl}) \qquad 1 \le k \le l$ and derive a new table 4.

Table4 customer information after the largest degree of membership

customer	basic call charge(A)	message charge(B)	GRPS charge(C)	Phone model(D)	the brand(E)
1	2	1	1	3	1
2	3	4	3	1	3
3	2	1	2	3	1
4	4	3	2	3	3
5	1	1	2	2	1
6	2	3	2	3	2
7	3	4	3	1	2
…	…	…	..	…	..

$$U/R = \{1,(2,7),3,4,5,6\},\ \ U/\{R-A\} = \{1,(2,7),3,(4,6),5\}$$
$$U/\{R-B\} = \{1,(2,7),(3,6),4,5\},\quad U/\{R-C\} = \{(1,3),(2,7),4,5,6\}$$
$$U/\{R-D\} = \{1,(2,7),3,4,5,6\},\quad U/S = \{(1,3,5),(2,4),(6,7)\};$$
$$\mu_R(S) = \{1,3,4,5,6\} + \{2,7\}i,\quad \mu_{R-A}(S) = \{1,3,5\} + \{2,4,6,7\}i \neq \mu_R(S)$$
$$\mu_{R-B}(S) = \{1,4,5\} + \{2,3,6,7\}i \neq \mu_R(S),\quad \mu_{R-C}(S) = \{4,5,6\} + \{1,2,3,7\}i \neq \mu_D(S)$$
$$\mu_{R-D}(S) = \{1,3,4,5,6\} + \{2,7\}i = \mu_R(S)$$

So the attribute D can be reduced to derive Table 5

Table5 customer information after attributes reduction

customer	basic call charge(A)	message charge(B)	GRPS charge(C)	the brand(E)
1	2	1	1	1
2	3	4	3	3
3	2	1	2	1
4	4	3	2	3
5	1	1	2	1
6	2	3	2	2
7	3	4	3	2
…	…	…	..	..

Through the investigation, we derive table 6 and table 7.

Table 6. Classical field of factors

the brand	basic call charge(Yuan/per month)	message charge	GRPS charge
Xin Shili	50-220	18-20	30-50
Shijie Feng	340-1000	2-5	2-5
Ruyi Tong	100-460	2-5	5-14

Table7. modulation field of factors

the brand	basic call charge(yuan/per month)	message charge	GRPS charge
Xin Shili	0-500	5-50	0-200
Shijie Feng	100-2000	0-50	0-500
Ruyi Tong	0-500	0-50	0-200

B. Decide right coefficient

When an object is evaluated, there is different degree in the standard. Right coefficient could be expressed the degree of the standard. Right coefficient has important effect on the evaluation result.

Therefore, there are four factors which are expressed from important to unimportant. The sequence is t age, basic call charge, message charge and GRPS charge. Then we decide right coefficient $\alpha = (0.36, 0.32, 0.32)$

C. Acquire object

According to the demand of the new customers, the china Unicom selects person A and B as objects. The value of each factor is acquired from two people. Then matter-element model of object is showed as follows.

$$M_A = \begin{bmatrix} O_A & basic\ call\ ch\arg e & 50 \\ & message\ ch\arg e & 20 \\ & GPRS\ ch\arg e & 20 \end{bmatrix}$$

Characteristic variables of objects are computed by using extension relevant function. The experiment result is expressed as Table 8.

Table8. The result of the relevant degree.

person	Xin Shili	Shijie Feng	Ruyi Tong	result
A	-0.16952	-0.43648	-0.32704	Xin Shili

From analysis in Table 8, it describes that person A belongs to the level of Xin Shili The analysis result can provide effective support for china Unicom to recommend an appropriate brand to new customers.

V. CONCLUSIONS

In the data base of information systems, there usually exists some attributes which are unimportant to decision attribute, people hope to find a minimum correlative attribute set which has the same capacity of classification for the decision attribute, the rules generate from the minimum attribute set is simpler and making more sense. And then extension transforming the condition and conductive transforming the result, acquire the variable knowledge, it is called extension knowledge. In fact, there are some bad records in data base that affect the quality of the data mining, making some rules not so reliable, so we must capture the main problem when we are solving the problem by extension transformation. This paper introduce selecting method of extension transforming attribute according the attributes' importance calculated during the process of the attribute reduction, this method guarantee the veracity of extension transformation, and improve the efficiency of the solution.

The result from traditional data mining is knowledge, but the result from extension data mining is strategy. Extension data mining has advantage in the decision-making of the enterprise. Extension relevant function can be used to quantitatively depict data gather of data mining. It can effectively solve the problem of the data quality, and provide clear decision-making for the enterprise. It also can provide a new thinking mode and approach for data mining which extenics is applied in. So extension data mining has great reference value.

REFERENCES:

[1]. Cai Wen. Extension theory and its application, Chinese Science Bulletin. 1999, 44(17):1538-1548

[2]. Cai Wen. Extension set and non-compatible problem, Advances Mathematics and Mechanics in China (in Chinese), Vol.2, Beijing: International Academic Publishers, 1990:1-21

[3]. Wen Cai. The extension set and incompatible problem. Science Exploration (in Chinese).1983, 3(1):83-97

[4]. Wen Cai. The Extension theory and its application. Chinese Science Bulletin. 1999, 44 (17):1538-1548

[5]. Wen Cai. "Matter-element Models and Their Application" (in Chinese), Science and Technology Documentation Publishers, Beijing, 1994.

[6]. Wen Cai, Chunyan Yang, and Weichu Lin. Methods of Extension Engineering, Science Press, Beijing, 1997

[7]. Chunyan Yan. Event element and its application. the theory of System Project and Practice,1998

[8]. Dongmei Zhu, Chunyan Yan. The extension of event element .the evolution of Artificial intelligence (in Chinese), the university of Beijing Post publisher, 2001.

[9]. http://web.gdut.edu.cn/%7Extenics

[10]. Huang Shiqiang, Conjugate system of matter-elements and the conjugate strategy. System Engineering Theory and Practice (in Chinese), 1998, 18(1): 117-122

[11]. Wang Wanliang, Zhao Yanwei. Research extension decision of mechanical intelligent CAD system. System Engineering Theory and Practice (in Chinese), 1998, 18(2):114-119

[12]. Zuo Jing.the extension domain of matter-elements based on the extension of matter-elements. Journal of Zhengzhou university of technology (in Chinese), Zhengzhou, 2001, 9(3): 22-27

[13]. YU Yongquan, ZHANG Jiwen. Generating available scheme for extension detecting technology [J]. China Engineering Science, 2001,4(1):64-68

[14]. Chen Wenwei. Research on mining the mutative knowledge with extension data mining. Engineering Science, 2006, 8(11):70-73

[15]. WenWei Chen, JinCai Huang .Extension Transformation and Extension Knowledge Representation of Attribute Reduction and Date Mining .Journal of Chongqing Institute of Technology ,2007:1-4

[16]. Li Lixi, Yang Chunyan, Li Huawen, Extension Strategy Generating System, Science Press, Beijing, 2006.

[17]. J.Han and M.Kamber. Data mining: concepts and techniques. Morgan Kaufman Publishers, San Francisco, CA. 2001.

[18]. PAWLAK Z. Why rough sets: proc.of the 5th IEEE International Conference on Fuzzy Systems[C]. New Orleans:IEEE,1996:738－743.

[19]. Wong SKM，Ziarko W. On optional decision rules in decision tables[J]．Bulletin of Polish Academy of Sciences, 33(11):693-696

[20]. Hu Keyun, Lu Yuchang, Shi Chunyi, "Advances in rough set theory and its applications". Journal of Tsinghua University (Sci & Tech), 41(1),2001,:64-68

[21]. Zhao Keqing. Set Pair Analysis and Its Primary Application [M]. Zhejiang Science Press, China, 2000.

Zhi-hang Tang (1974-), male, born in hunan province China, School of Computer and Communication, Hunan Institute of Engineering, Doctor, research field: intelligent Decision and data mining.