Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXTENSION DATA MINING BASED ON EXTENSION SET AND ROUGH SET

[1,2]ZHI-HANG TANG, [2]BEI-PING TANG

[1] Assoc. Prof., School of Computer and Communication, Hunan Institute of Engineering Xiangtan 411104, China

[2] Assoc. Prof., College of Business Administer, South China University of Technology Guangzhou 510640, China

E-mail: [1]tang83769@126.com, [2]tang2018@sina.com.

ABSTRACT

Data mining has gained popularity in the database field recently, the gained knowledge is static because of the static nature of the database, and does not reflect the dynamic nature of knowledge. Extension data mining is a product combining Extenics with data mining, By using the theory and method of Extenics, it can mine the knowledge from database which is relative to solve contradictory problems. And the knowledge includes the Extension classification knowledge, conductive knowledge and other knowledge associated with transformation, which collectively called Extension knowledge. Its Application to enterprise brands segmentation will help the enterprises to gain their ends on the best way. Research result indicates that extension data mining can provide effective decide support for the Decision-making of enterprise.

Keywords: *Extension Data Mining (EDM), Attributes Reduction, Rough Set*

1. INTRODUCTION

Data mining [1] 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 [2-5]. Some of these techniques are beginning to be scaled to operate on databases.

It is a kind of automated data analysis techniques based on databases or data warehouses. It can quickly find new knowledge valuable, meaningful and with specific relationships in vast amounts of data. These new knowledge can provide managers with effective decision support. But there are still some problems. For example, the gained knowledge is static because of the static nature of the database, and does not reflect the dynamic nature of knowledge.

The human history is a history about solving contradictory problem. People are constantly dealing with the contradict problem in the past, the present and the future. The tool to solve the contradictory problem is transformation. Therefore, the research about extension transformations is the important part of solving the contradictory problem. In the modern society, all trades and professions have accumulated vast amounts of data. Now the question is: how can

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

we mine the transformation knowledge from these data to assist us with solving the contradictory problem? And this question proposes a new subject for data mining.

Extenics[6] 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[7].Extension data mining [8] (EDM) is a product combining Extenics with data mining. Extension data mining was proposed in 2004. After several years' researching and exploring, we finally have clarified the object and objective of this study. By using the theory and method of Extenics, it can mine the knowledge from database which is relative to solve contradictory problems. And the knowledge includes the Extension classification knowledge, conductive knowledge and other knowledge associated with transformation, which collectively called Extension knowledge. At present, we preliminary explored some questions of EDM, just the basic theory, basic method and their implementation on computers.

We will go forward to information overload era from data explosion and knowledge explosion era. In the knowledge overload era, businesses' decision-maker is more need to be supported by practical knowledge. For example, in an increasingly competitive market, customers have become important resources today. The transformation knowledge will help the initial registration users and customers who will soon leave to turn into the loyal customer, so as to reduce the cost of customer retention and new customer development. During credit risk analyzing, it is not only necessary to identify high-risk customers, but also have to take measures to stop fraud actions of motivated customers, then the classification methods and their extension-related knowledge will be on the great development potentials. During new products development, implicational analysis helps to find product trends earlier and identify the potential need of customer. During business process optimizing, extension data mining can contribute to identify bottlenecks in the efficiency and take transformation measures. In the medical industry, transformed knowledge can help doctors to detect fundamental change symptoms much earlier and identify the most effective programs to improve treatment. In marketing, the transformation knowledge has a guiding significance to market development. All in all, extension data mining can play a role in things classification transformation, finding the root causes of the problem, identifying potential transformation knowledge, and so on. Therefore, extension data mining has broad application prospects.

2. BASIC CONCEPTION

2.1 Rough Sets

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[9], 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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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[10] 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[11].

2.2 The Set Theory Foundation Of EDM

The set is a mathematical method of classifying and identifying the objective things by our brain. At the end of 19th century, a Germen named Cantor propounded Cantor Set, which is used to classify the confirmatory things. The Cantor set uses characteristic function valued by {0, 1} to characterize whether object belongs to the set. Cantor Set cannot describe the fuzziness of things and fuzzy things. In 1965, an American named Zadeh propounded the concept of Fuzzy Set, which can characterize the degree of having a certain property by using function valued by the interval [0, 1]. However, in numerous problems, the degree of having a certain property is changeable. During the problem solving, things with a certain property change into without, and problem with contradiction change into without. In order to describe the positive and negative in a certain condition, the concept of Extension Set was pronounced, which describes the degree of something having a certain property by using function valued by the interval (-∞ , + ∞). Extension Set also tries to use Extension Domain to turn something with a certain property to without. In other words, Extension Set is a kind of changeable Set which describes the variability of things and researches variable classification. EDM mines the changeable knowledge by using the expanding minds of Extension Set.

Extension set is the set theory foundation to study the variable classification. It has a wide range of applications, such as market segmentation, enterprise customer classification, product classification, the customer value research.

2.3 The Main Contents Of EDM

Extension Data Mining (EDM) focuses on mining extension knowledge based on extension transformations. At present, the mining objects of EDM are mainly relational database or data warehouse, which is typical structured data. With in-depth research and technology development, the research objects will be gradually extended to the semi-structured or unstructured data, such as text data, image, video, data and Web data.

EDM matches the data with elements, and matches the database and data warehouse with the domain of extension set. EDM manage to apply the extension set theory and extension logic into data mining in theory, consequently form the basic theory of mining extension knowledge. In the method, EDM manage to match database, data warehouse and formal system which built on element logic cell, and form a knowledge representation method which is suitable for data transformation. Using extension reasoning and extension tools such as correlation function, we can establish extension data mining methods suitable for mining extension knowledge.

EMD adopts extension classification method, expansion analysis method, extension transformation method, extension reasoning

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

method and excellent degree evaluation method of Extenics. On the other hand, it calculates correlation degree of information element quantitatively with correlation function. In other words, a combination of qualitative and quantitative is the basic principles of extension data mining. In addition, except for mining extension knowledge from the database directly, EDM try to research new mining methods based on the existing rule-based knowledge of knowledge base. Currently, the extension data mining research mainly focus on extension classification knowledge mining, conductive knowledge mining and extension clustering knowledge mining which is based on database, the extension knowledge mining based on knowledge base and the computer implementation of these methods with the cases.

3. EXTENSION DATA MINING BASED ON ROUGH SET

3.1 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.

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

$$Core(B) = \bigcap Red(B) \quad (1)$$

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.

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

$$CORE^x(B) = \bigcap Red^x(B) \quad (2)$$

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

Suppose we are given a dependency $C \Rightarrow D$. 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 $C,D \subseteq A$. Obviously if $C' \subseteq C$ is a D-reduct of C, then C' is a minimal subset of C such that

$$\gamma(C,D) = \gamma(C',D) \quad (3)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- 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

$$CORE_D(C) = \bigcap Red_D(C) \quad (4)$$

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

If D = C we will get the previous definitions.

$$IND(R) = IND(R - \{r\}) \quad (5)$$

Where ind ( ) denotes the indiscernibility relation, and r ∈R, which is the attribute sets. Obviously, if Eq. (5) 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.

3.2 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.

3.3 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.

$$M_{cf} = (O_{cf}, c, v) = \begin{bmatrix} O_{cf} & C_1 & \langle v_{cf1}^l, v_{cf1}^h \rangle \\ & C_2 & \langle v_{cf2}^l, v_{cf2}^h \rangle \\ & \vdots & \vdots \\ & C_n & \langle v_{cfn}^l, v_{cfn}^h \rangle \end{bmatrix} \quad (6)$$

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 $\langle v_{cfi}^l, v_{cfi}^h \rangle (i = 1,2,\cdots,n)$ which is a classical field. This is similar to $X_0 = \langle a, b \rangle$.

$$M_{mf} = (O_{mf}, c, v) = \begin{bmatrix} O_{mf} & C_1 & \langle v_{mf1}^l, v_{mf1}^h \rangle \\ & C_2 & \langle v_{mf2}^l, v_{mf2}^h \rangle \\ & \vdots & \vdots \\ & C_n & \langle v_{mfn}^l, v_{mfn}^h \rangle \end{bmatrix} \quad (7)$$

v_{mf} is the variable range which is ensured by characteristic variables of O_{mf} .So v_{mf} is called

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$\langle v_{mfi}^{l}, v_{mfi}^{h} \rangle (i = 1,2,\cdots,n)$ which is a modulation field. This is similar to $X = \langle c, d \rangle$.

3.4 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:

$$k_{ij} = \sum_{i=1}^{n} \alpha_{ij} \cdot \frac{k_j(x_i)}{\max|k_j(x_i)|}, \quad i = 1,2,\cdots n; j = 1,2,\cdots m \quad (8)$$

$$k_i = \bigvee_{j=1}^{m} k_{ij} \quad (9)$$

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: $\alpha_{ij} = \frac{x_j / b_{ij}}{\sum_{j=1}^{m} x_j / b_{ij}}$ (10)

Finally, we determine the category of the testing sample.

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

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

4. CONCLUSIONS

With the rapid development of information technology, management information systems, Internet, data mining and knowledge management are accumulating more and more of data, information and knowledge. We will go forward to information overload era from data explosion and knowledge explosion era. In the knowledge overload era, businesses' decision-maker is more need to be supported by practical knowledge. For example, in an increasingly competitive market, customers have become important resources today. The transformation knowledge will help the initial registration users and customers who will soon leave to turn into the loyal customer, so as to reduce the cost of customer retention and new customer development. During credit risk analyzing, it is not only necessary to identify high-risk customers, but also have to take measures to stop fraud actions of motivated customers, then the classification methods and their extension-related knowledge will be on the great development potentials. During new products development, implicational analysis helps to find product trends earlier and identify the potential need of customer. During business process optimizing, extension data mining can contribute to identify bottlenecks in the efficiency and take transformation measures. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the medical industry, transformed knowledge can help doctors to detect fundamental change symptoms much earlier and identify the most effective programs to improve treatment. In marketing, the transformation knowledge has a guiding significance to market development. All in all, extension data mining can play a role in things classification transformation, finding the root causes of the problem, identifying potential transformation knowledge, extension data mining has broad application prospects.

ACKNOWLEDGEMENTS

This work is supported by 2012 project of education department of hunan province (project number: **12C0628 and 12C0639**) and China Postdoctoral Science Foundation (Grant no. 20110490888).

REFRENCES

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

[2]. Wu Roung-Shiunn,Yen, David C.Using data mining technique to enhance tax evasion detection performance.*Expert Systems with Applications*, Vol. 39. No. 10, 2012, pp.8769-8777

[3]. Morik, Katharina, Bhaduri, Kanishka,Kargupta, Hillol.Introduction to data mining for sustainability.*Data Mining and Knowledge Discovery*, Vol. 24. No. 2, 2012, pp.311-324

[4]. MANOHAR ANNAPPA KOLI, S BALAJI, EXTRACTION OF BINARY PATTERNS FOR IMAGE DE-NOISING USING DATA MINING, *Journal of Theoretical and Applied Information Technology*, Vol. 46. No. 1, 2012, pp.037- 045

[5]. JING TAO, YUAN YIN, "INTELLIGENT DESIGN SYSTEM OF MECHANICAL PRODUCTS BASED ON DATA MINING AND KNOWLEDGE BASED ENGINEERING", *Journal of Theoretical and Applied Information Technology*, Vol. 46. No. 1, 2012, pp.237-244

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

[7]. Wang Wanliang, Zhao Yanwei. Research extension decision of mechanical intelligent CAD system. *System Engineering Theory and Practice* (in Chinese), Vol. 18. No. 2, 1998, pp.114-119

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

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

[10]. Wong SKM，Ziarko W. On optional decision rules in decision tables[J]. *Bulletin of Polish Academy of Sciences*, Vol. 33. No. 11,2011, pp.693-696

[11]. Hu Keyun, Lu Yuchang, Shi Chunyi, "Advances in rough set theory and its applications". *Journal of Tsinghua University (Sci & Tech)*, Vol. 41. No. 1, 2001,pp.64-68