Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06), 457-466, 2006.

User-centered Interactive Data Mining

Yan Zhao, Yaohua Chen and Yiyu Yao
Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{*yanzhao, chen115y, yyao*}*@cs.uregina.ca*

Abstract

While many data mining models concentrate on automation and efficiency, interactive data mining models focus on adaptive and effective communications between human users and computer systems. User views, preferences, strategies and judgements play the most important roles in human-machine interactivities, guide the selection of target knowledge representations, operations, and measurements. Practically, user views, preferences and judgements also decide strategies of abnormal situation handling, and explanations of mined patterns. In this paper, we discuss these fundamental issues.

1. Introduction

Exploring and extracting knowledge from data is one of the fundamental problems in science. Many methods have been proposed and extensively studied, such as database management, statistics, machine learning, etc. Particularly, data mining takes up many important tasks, such as description, prediction and explanation of data.

Data mining is featured by applying computer technologies to carry out nontrivial calculations. Computer systems can maintain precise operations under heavy information load, and maintain steady performance. Without the aid of computer systems, it is very difficult for people to aware, extract, memorize, search and retrieve knowledge in large and separate datasets, to interpret and evaluate data and information that are constantly changing, to make recommendations or predictions in the face of inconsistent and incomplete data.

It is true that computer technologies have freed humans from many time-consuming and labor-intensive activities. However, full automation of cognitive functions such as decision making, planning, and creative thinking remains human's job. Implementations and applications of computer systems reflect requests and preferences of human users, and contain certain human heuristics. Computer systems must rely on human users to set goals, select alternatives if original approach fails, participate in unanticipated emergencies and novel situations, and develop innovations in order to preserve safety, avoid expensive failure, or increase product quality [11, 16, 29].

According to the above observations, we believe that interactive systems are required for data mining tasks. Though human-machine interaction has been emphasized for many disciplines, it did not get enough attention in the domain of data mining until recently [3, 15, 50]. Generally, an interactive data mining system is an integration of a human user and a computer. They can communicate and exchange information and knowledge. A foundation of human-computer interaction may be provided by cognitive informatics [36, 37, 38].

Through interaction and communication, computers and users can divide the labors in order to achieve a good balance of automation and human control. Computers are used to retrieve and keep track of large volumes of data, and to carry out complex mathematical or logical operations. Users can avoid routine, tedious, and error-prone tasks, concentrate on critical decisions, planning, and cope with unexpected situations [11, 29]. Moreover, interactive data mining can encourage learning, improve insights and understandings of the domain, stimulate the exploration of creative possibilities, and help users to solve particular problems. Users' feedback can be used to improve the system. The interaction is bi-beneficial.

In this paper, we discuss some of the fundamental issues of user-centered interactive data mining. It is important to note that users possess various skills, intelligence, cognitive styles, frustration tolerances and other mental abilities. They come to a problem with various preferences, requirements and background knowledge. Given a set of data, every user may try to make sense out of data by seeing it from different angles, in different aspects, and under different views. Based on these differences, data mining methods and results are not unique. There does not exist a universally applicable theory or method to serve the needs of all users. This motivates and justifies the co-existence of many theories and methods for data mining systems, as well as the exploration of new theories and methods.

In Section 2, we talk about multiple views of data mining and knowledge discovery. We exemplify various measures that associate and evaluate multiple views. Interactive data mining systems should support and encourage multiple views. It is often meaningless to argue which view is better, more suitable or more appropriate by isolating it from user requirements and applications.

While a specific view is targeted, some standards and efficient approaches can be used or implemented. At this stage, user-centered interaction may be characterized as a user preference. A user may prefer one solution to another, one arrangement to another, one attribute order to another, or, one result to another. In Section 3, we present a formal model of user preference. Based on the model, a user preference is represented as a weak order.

Having a specific view being targeted and a user preference being provided, an interactive data mining system can carry out calculations and inferences. However, in many real applications, ideal situations do not exist. Instead, users need to choose a strategy for abnormal situation handling. In Section 4, we discuss several different strategies, each represents a particular type of user requirements. Some of the abnormal situation handlers have been embedded into specific algorithms. We examine how these handlers are different from each other.

In the existing data mining frameworks, the phase of interpretations and explanations of the discovered knowledge is often missing. In Section 5, we talk about how to explain a discovered pattern, in order to relate it more closely to a user.

As a whole, multiple views, user preferences, multiple strategies and explanation construction are all user-centered. They together form our understanding of interactive data mining.

2. Multiple Views in Data Mining

Many techniques and models of machine learning and statistics can be applied to data mining. Typically, each model presents a particular and single view of data, or discovers a specific type of knowledge embedded in data. It explores different types of knowledge, different features of data, different user requirements, and different interpretations of data [5]. Multiple views imply that one is able to derive multiple hierarchies for the same system [44]. Each hierarchy is defined based on a particular interpretation, and can be broken down into multiple levels. With respect to a particular hierarchy, levels represent localized views, and they are tied together to form a global view. One may climb up and down the hierarchy to study a system at various resolution levels.

2.1. Information tables and the logic language

An information table represents all available information. Knowledge or rules can be discovered based on information tables. The rows of an information table represent the objects. The columns describe a set of attributes. An information table can be formally defined by a quadruple:

$$S = (U, At, \{V_a \mid a \in At\}, \{I_a \mid a \in At\}),$$

where U is a finite nonempty set of objects, At is a finite nonempty set of attributes, V_a is a nonempty set of values for $a \in At$, and I_a is an information function $I_a : U \rightarrow V_a$ for $a \in At$.

With respect to the definition of information table, a logic language can be defined to express various types of rules. We adopt the decision logic language studied by Pawlak [25]. Similar languages have been studied by many authors [8, 41]. In this language, an atomic formula is a pair (a, v), where $a \in At$ and $v \in V_a$. If ϕ and ψ are formulas, then so are $\neg\phi, \phi \wedge \psi, \phi \vee \psi, \phi \rightarrow \psi$, and $\phi \leftrightarrow \psi$ by applying the logic connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$. For an atomic formula $\phi = (a, v)$, it is assumed that an object x satisfies ϕ or does not satisfy ϕ. In other words, if an object x has a value v on the attribute a, then we say that the object x satisfies ϕ. Otherwise, we say x does not satisfy ϕ. The set of objects that satisfy a formula ϕ is denoted as $m(\phi)$. For an atomic formula is a pair (a, v), the set of objects is $m(a, v) = \{x \in U \mid I_a(x) = v\}$. The following properties hold:

$$\begin{array}{lllll}
\text{(i)} & m(\neg\phi) & = & \neg m(\phi), \\
\text{(ii)} & m(\phi \wedge \psi) & = & m(\phi) \cap m(\psi), \\
\text{(iii)} & m(\phi \vee \psi) & = & m(\phi) \cup m(\psi), \\
\text{(iv)} & m(\phi \rightarrow \psi) & = & \neg m(\phi) \cup m(\psi), \\
\text{(v)} & m(\phi \leftrightarrow \psi) & = & (m(\phi) \cap m(\psi)) \cup \\
 & & & (\neg m(\phi) \cap \neg m(\psi)).
\end{array}$$

The formula ϕ can be viewed as the description of a set of objects in $m(\phi)$.

In the study of formal concepts, every concept consists of two parts, the intention and extension [14, 39]. A formula ϕ represents the intention of a concept and a set of objects $m(\phi)$ denotes the extension of the concept. The pair $(\phi, m(\phi))$ is denoted as a concept, which can be described as the set of objects $m(\phi)$ that having the feature expressed by the formula ϕ.

2.2. Classes of rules

Knowledge generated from a large data set is often expressed in terms of a set of discovered rules. One of the important tasks of data mining is to find strong relationships between concepts [41]. A rule can be defined and represented as $\phi \Rightarrow \psi$, where ϕ and ψ are intentions of two concepts. The symbol $\Rightarrow$ represents a connection or a relationship between two concepts ϕ and ψ. The meanings and interpretations of $\Rightarrow$ are varied based on user requirements or applications. Rules can be classified according to the interpretations of $\Rightarrow$. In other words, different interpretations of $\Rightarrow$ in a rule represent different types of knowledge. It might be impossible to list all classes of rules. We only discuss several of them.

An association rule $\phi \Rightarrow \psi$ describes an association relationship between two concepts ϕ and ψ. That is, when ϕ occurs, ψ occurs too.

A classification rule $\phi \Rightarrow \psi$ presents a decision relationship between the two concepts. That is, if ϕ occurs then ψ occurs.

Sometimes, users prefer to use a subset of attributes $A \subseteq At$ to define a rule. The benefits of such attribute selected rules are that they are normally shorter, cheaper, and hence more understandable, actionable and profitable. The attribute set A contains attributes that singly necessary and jointly sufficient to keep the information provided by the original attribute set. In the term of rough set theory, A is called a *reduct* [24, 25].

In other situations, the attribute set A defines a unique object set $U' \subseteq U$, such that all the objects possess A are in U', and A is the attribute set including all the features shared by U'. Such a bounding relationship between A and U' can be explained by formal concepts in the domain of formal concept analysis [39, 40].

2.3. Objective and subjective measures

Based on the extensions, various quantitative measures can be used for rule interestingness evaluation [45, 49].

Measures can be classified into two categories: objective measures and subjective measures [30]. Objective measures depend on the structure of rules and the underlying data used in the discovery process. Subjective measures depend on the user beliefs [22, 30].

Measures defined by statistical and structural information are viewed as objective measures. For example, Gago and Bento proposed a measure for selecting discovered rules with the highest average distance between them [13]. The distance measure is developed based on the structural and statistical information of a rule such as the number of attributes in a rule and the values of attributes. A rule is deemed as interesting if it has the highest average distance to the others. One does not consider the application and domain when measuring the discovered rules by using the distance measure. Information theoretic measures are also objective measures because they use the underlying data in a data set to evaluate the information content or entropy of a rule [20, 32, 42].

Different types of rules can be identified based on different objective measures. For example, peculiarity rules have low support and high confidence, exception rules have low support and high confidence, but complement to other high support and high confidence rules, and outlier patterns are the ones that far away from the statistical mean [49, 51, 52].

Although statistical and structural information provides an effective indicator of the potential effectiveness of a rule, its usefulness is limited. One needs to consider the subjective aspects of rules. Subjective measures consider the user who examines the rules. For example, Silberschatz and Tuzhilin proposed a subjective measure of rule interestingness based on the notion of unexpectedness and in terms of a user belief system [30, 31]. The basic idea of their measure is that the discovered rules which have more unexpected information with respect to a user belief system are deemed as more interesting. Thus, subjective measures are both application and user dependent. In other words, a user needs to incorporate other domain specific knowledge such as user interest, utility, value, profit, action-ability, etc. [27, 34].

As one example, profit or utility-based mining is a special kind of constraint-based mining, taking into account of both statistical significance and profit significance [35]. Doyle discussed the importance and usefulness of the notions of economic rationality and suggested that economic rationality can play a large role for measuring a rule [10]. Similarly, Barber and Hamilton proposed the notion of share measures which consider the contribution, in terms of profit, of an item in an item set [2].

2.4. Association measures of concepts

Many quantitative measures are proposed to evaluate different relationships between attribute sets. Each measure reflects a different and specific feature of data. We consider several types of measures below.

One-way association

Confidence is a commonly used measure for evaluating association of a rule. The basic idea of confidence is described as the probability that concept ψ occurs given that concept ϕ occurs [1]:

$$\begin{aligned} P(\phi \mid \psi) &= \frac{|m(\phi) \cap m(\psi)|}{|m(\psi)|} \\ &= \frac{P(\phi, \psi)}{P(\phi)}, \end{aligned}$$

where $P(\phi)$ is called the *support* of ϕ and is defined by $P(\phi) = \frac{|m(\phi)|}{|U|}$.

Confidence is one-direction from ϕ to ψ and can be viewed as a one-way association measure [49]. In other

Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06), 457-466, 2006.

words, the concept ϕ depends on the concept ψ, but ψ may not depend on ϕ.

Two concepts ϕ and ψ are viewed as being non-associative or independent if the occurrence of ϕ does not alter the probability of ψ occurring. In other words, if the occurrence of ϕ can affect the probability of ψ, then we say that the concept ψ is dependent on or associated with the concept ϕ. Typically, a rule with a high support and a high confidence are considered having a strong association relationship between two concepts.

Two-way association

RI is also a measure on the evaluation of the association of a discovered rule [26]. It is defined by:

$$RI \quad = \quad P(\phi, \psi) - P(\phi)P(\psi).$$

The two concepts ϕ and ψ are recognized as being non-associative or independent when $RI = 0$ ($P(\phi)P(\psi) = P(\phi, \psi)$). In fact, this measure determines the degree of association of a rule by the comparison of the joint probability of two concepts $P(\phi, \psi)$ with respect to the expected probability of the non-association assumption $P(\phi)P(\psi)$. $RI > 0$ represents a positive association from ϕ to ψ. $RI < 0$ represents a negative association, which is from ψ to ϕ.

IND is similar to the measure of rule-interest [4]. This measure is defined by:

$$IND = \frac{P(\phi, \psi)}{P(\phi)P(\psi)}.$$

The two concepts ϕ and ψ are recognized as being non-associative or independent when $IND = 1$ ($P(\phi)P(\psi) = P(\phi, \psi)$). This measure is the ratio of the joint probability of $\phi \cap \psi$ and the probability obtained if ϕ and ψ are assumed to be independent. In other words, the rule has a stronger association if the joint probability is further away from the probability under independence.

The *IS* measure is another similar measure with respect to the measure of rule-interest [33]. It can be defined by:

$$\begin{aligned} IS \quad &= \quad \frac{P(\phi, \psi)}{\sqrt{P(\phi)P(\psi)}} \\ &= \quad \sqrt{P(\phi \mid \psi)P(\psi \mid \phi)}. \end{aligned}$$

The basic notion of the *IS* measure is similar to the measure of independence. Furthermore, it is equivalent to the geometric mean of confidences of the rule. However, its range is between 0 and 1 instead of *IND*'s range, between 0 and ∞.

The *RI*, *IND*, and *IS* measures are symmetric and viewed as two-way association measures [49]. If two concepts ϕ and ψ in a rule have a two-way association relationship, then the concept ϕ must depend on or be associated with the concept ψ, and the converse is also true.

2.5. Correlation measures of attributes

The association relation we discussed above shows the relationship between two concept intentions, each is defined by a specific combination of one attribute and one of its possible values. This is also called a *local connection*, generates a *low order* rule. A *global connection* is characterized by showing the relationships between all combinations of values on one set of attributes and all combinations of values on another set of attributes. It is also called a *high order* rule, revealing the correlation of two attribute sets [43].

A statistical measure, called *correlation coefficient*, can be used to compute the degree of correlation between two numerical attributes $X, Y \in At$ [33]:

$$r(X,Y) = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}},$$

where $SS_X = \sum(x - \overline{x})^2$, $x \in V_X$, $\overline{x}$ is the mean value of x. $SS_Y = \sum(y - \overline{y})^2$, $y \in V_Y$, $\overline{y}$ is the mean value of y. $SS_{XY} = \sum(x - \overline{x})(y - \overline{y})$ is the covariance between the attributes X and Y. $\sqrt{SS_X}$ and $\sqrt{SS_Y}$ are the standard deviations of the real values with respect to the mean value. By extending this equation, we have:

$$r(X,Y) \quad = \quad \frac{E(X \wedge Y) - E(X)E(Y)}{\sigma_X \sigma_Y},$$

where $E(X \wedge Y) = n \sum xy$, $E(X) = \sum x$, and $E(Y) = \sum y$ are the expected values on the attributes $X \wedge Y$, X, and Y respectively. σ_X and σ_Y denote the standard deviations.

For a discovered high order rule $X \Rightarrow Y$, we suppose that there exist two value sets V_X and V_Y in an information table for the two attributes. If V_Y increases or decreases as V_X increases, then the two attributes are considered as being closely correlated.

The correlation coefficient is a number between 0 and 1. The closer the correlation coefficient to 1, the stronger the correlation between attributes. If the two attributes are not correlated, $r(X,Y)$ is zero because of $SS_{XY} = 0$ and $E(X,Y) = E(X)E(Y)$. In other words, if X and Y are not closely related to each other, they do not "co-vary", the covariance is small and the correlation is small. If X and Y are closely related, the covariance is almost the same as $\sigma_X \cdot \sigma_Y$ and the correlation is almost 1.

Correlated attributes may not necessarily be dependent or associated. Also, attributes that are associated may not necessarily be correlated. Mari and Kotz analyzed several common and different features of association and correlation measures from the statistical point of view [23]. Correlation coefficient only evaluates the linear relationship between attributes, but there are situations in which linear correlation do not exist but a strong nonlinear association exists between attributes.

Brin *et al.* proposed the use of Chi-square (χ^2) probability testing to evaluate the association between two attributes in a discovered rule [4]. If the Chi-squared value is 0, then the attributes are independent, otherwise, they are dependent on each other. However, the χ^2 testing does not give the strength of the association between attributes in a discovered rule [33]. Instead, it can only decide whether attributes in a rule are non-associative or associative. Therefore, it cannot be used to rank the discovered rules. Liu *et al.* suggested to prune the insignificant rules by using the standard χ^2 test combined with a support-confidence test [21].

2.6. Single rule measures and multiple rule measures

The measures of rule interestingness can also be classified into measures for a single rule and measures for a set of rules. Furthermore, a measure for a set of rules can be obtained from measures for single rules. For example, conditional probability can be used as a measure for a single classification rule, and conditional entropy, which is defined by conditional probability, can be used as a measure for a set of classification rules [42].

Measures for multiple rules concentrate on features of a set of rules. They are normally expressed as some kinds of average. Many measures, known as *summaries*, have been examined for multiple rules [17].

One may consider integrated and general systems to satisfy the various user requirements. Interactive approaches can be viewed as one of the potential solutions.

3. User Preferences

A user preference makes relation between a user and a target item which contains several features. An item can be a target concept to be learned, a system, a model, an algorithm, or an approach that is ready to be chosen. A feature is normally associated with feature description, measures of this feature, and a set of possible feature values, or feature value ranges. Usually preference of an item is indirectly related to the preference of its features [18]. In this section, we discuss different types of user preferences.

3.1. User preference modelling

User judgement can be expressed in various forms. Quantitative judgement involves the assignment of different weights to different items. Qualitative judgement is expressed as an ordering of items. In many situations, user judgement is determined by semantic considerations. For example, it may be interpreted in terms of more intuitive notions, such as the cost of testing, the easiness of understanding, or the actionability. It is virtually impossible to list all practical interpretations of user judgement. In addition, the meaning of user judgement becomes clear only in a particular context of application.

Quantitative user preferences

A simple and straightforward way to represent user judgement on items is to assign them with numerical weights [19]. Formally, it can be described by a mapping:

$$w : X \longrightarrow \Re, \tag{1}$$

where X is a finite non-empty set of items, and $\Re$ is the set of real numbers. For an item $a \in X$, $w(a)$ is the weight of a. The numerical weight $w(a)$ may be interpreted as the degree of importance of a, the number of occurrences of a in a set, or the cost of testing a in a rule. This induces naturally an ordering of items.

Qualitative user preferences

A difficulty with the quantitative method is the acquisition of the precise and accurate weights of all items. On the other hand, a qualitative method only relies on pairwise comparisons of items. For any two items $a, b \in X$, we assume that a user is able to state whether one is more important than, or more preferred to, the other. This qualitative user judgement can be formally defined by a binary preference relation $\succ$ on X. For any two $a, b \in X$:

$$a \succ b \Longleftrightarrow \text{ the user prefers } a \text{ to } b. \tag{2}$$

In the absence of preference, i.e., if both $\neg(a \succ b)$ and $\neg(b \succ a)$ hold, we say that a and b are indifferent. An indifference relation $\sim$ on X is defined as:

$$a \sim b \Longleftrightarrow \neg(a \succ b) \wedge \neg(b \succ a). \tag{3}$$

Based on the strict preference and indifference, one can define a preference-indifference relation $\succeq$ on X:

$$a \succeq b \Longleftrightarrow a \succ b \vee a \sim b. \tag{4}$$

If $a \succeq b$ holds, we say that b is not preferred to a, or a is at least as good as b. The strict preference can be re-expressed as $a \succ b \Longleftrightarrow a \succeq b \wedge \neg(b \succeq a)$.

A user preference relation satisfies two axioms: asymmetry and negative transitivity, so it is a *weak order* on X. For any $a, b, c \in X$:

$$a \succ b \Longrightarrow \neg(b \succ a);$$
$$(\neg(a \succ b) \wedge \neg(b \succ c)) \Longrightarrow \neg(a \succ c).$$

The asymmetry axiom states that a user cannot prefer a to b, and at the same time prefer b to a. The negative transitivity axiom states that if a user does not prefer a to b, nor b to c, then the user should not prefer a to c.

Proceedings of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'06), 457-466, 2006.

A weak order imposes a special structure on the set X of items. The indifference relation $\sim$ divides the set of items into disjoint subsets. Furthermore, for any two distinct equivalence classes $[a]_\sim$ and $[b]_\sim$ of $X/\sim$, either $[a]_\sim \succ' [b]_\sim$ or $[b]_\sim \succ' [a]_\sim$ holds. In other words, it is possible to arrange the items into several levels so that items in a higher level are preferred to items in a lower level, and items in the same level are indifferent.

When each equivalence class contains exactly one item, the preference relation $\succ$ on X is in fact a linear order itself. In general, if we do not care how to order items in an equivalence class, we can extend a weak order into a linear order such that a is ranked ahead of b if and only if $a \succeq b$. For a weak order, its linear extension may not be unique [12].

Connections of quantitative and qualitative preferences

The quantitative judgement can be easily translated into qualitative judgement. Given the weights of items, we can uniquely determine a preference relation. Suppose there are two items a and b, $w(a)$ and $w(b)$ represent the importance of a and b, respectively, a preference relation is defined by:

$$a \succ b \Longleftrightarrow w(a) > w(b). \tag{5}$$

When $w(a)$ and $w(b)$ is the cost of items a and b, the following preference relation should be used instead,

$$a \succ b \Longleftrightarrow w(a) < w(b). \tag{6}$$

In general, two items may have the same weights. The induced preference relation is indeed a weak order, i.e., asymmetric and negatively transitive.

The translation to a preference relation only preserves the ordering of items implied by the weights. The additional information given by the absolute weight values is lost. In the reverse process, a user preference relation can be represented in terms of the weights of items. A rational user's judgement must allow numerical measurement.

The following theorem states that a weak order is both necessary and sufficient for a numerical measurement [12, 28]:

Theorem 1 *Suppose $\succ$ is a preference relation on a finite non-empty set X of items. There exists a real-valued function $u : X \Rightarrow \Re$ satisfying the condition:*

$$a \succ b \Longleftrightarrow u(a) > u(b), a, b \in X, \tag{7}$$

if and only if $\succ$ is a weak order. Moreover, u is uniquely defined up to a strictly monotonic increasing transformation.

The function u is referred to as an order-preserving utility function. It provides a quantitative representation of a user preference. That is, the numbers of $u(a), u(b), \ldots$ as ordered by $>$ reflect the order of $a, b, \ldots$ under the preference relation $\succ$.

The utility function also trustfully represents the indifference relation, that is,

$$a \sim b \Longleftrightarrow u(a) = u(b), a, b \in X. \tag{8}$$

According to Theorem 1, for a given preference relation, there exist many utility functions. The utility functions are in fact based on the ordinal scale. That is, it is only meaningful to examine the order induced by a utility function. Although numerical values are used, it is not necessarily meaningful to apply arithmetic operations on them.

4. Multiple Strategies for Abnormal Situation Handling

In solving real world problems, we often face the choices between simple and complicated descriptions, precise and imprecise characterizations, understandability and incomprehensibility of methods, and exact and approximate solutions. In general, there is a tradeoff of such two opposite criteria of the competing nature. Human problem solving depends crucially on a proper balance and compromise of these incompatible criteria. Different users can develop different knowledge representation frameworks and related automated learning and mining mechanisms to describe and identify abnormal situations or behaviors. Consequently, this issue must be addressed in the user-centered interactive data mining.

4.1. Retaining strategies

A retaining strategy, by its name, means to keep the quality of the rules, especially their accuracy, as high as they could be. The most commonly used accuracy measure is the confidence measure defined in the last section. Clearly, the higher the confidence value, the more accurate the rule is. In most real situations, a rule, in the form of $\phi \Rightarrow \psi$, is not always deterministic for the given universe, but rather approximate and uncertain. In other words, the confidence value of the rule is less than or equal to 100%.

Yao *et al.* advocated the use of a specific knowledge representation and data mining framework based on rules and exceptions [47]. In this framework, normal and abnormal situations or behaviors occur as pairs of dual entities: rule succinctly summarizes normal situations, and exceptions characterize abnormal situations. These two entry types each provides the context for defining the other. Rule+exception strategies strike a practical balance between simplicity and accuracy.

Two types of exceptions can be identified, incorrect interpretations produced by the existing rules, and the interpretations that cannot be produced by the existing rules [6]. For simplicity, they are refereed to as incorrectly covered exceptions and uncovered exceptions, respectively.

For the incorrectly covered exceptions, two potential solutions exist. A commonly used method adds an additional condition ϕ' to form a more specific condition $\phi \wedge \phi'$. The new rule $\phi \wedge \phi' \Rightarrow \psi$ should produce fewer or no exceptions. Another alternative, a rule+exception strategy treats $\phi \Rightarrow \psi$ as a general rule with probability and searches for exception rules or exception instances to the general rule.

For uncovered exceptions, we could attempt to add an alternative condition to form a more general rule $\phi \vee \phi' \Rightarrow \psi$. The extra ϕ' could cover more instance of ψ. For clarity, we can think of them as two rules $\phi \Rightarrow \psi$ and $\phi' \Rightarrow \psi$. One can view the second rule as an exception rule to handle the uncovered exceptions. In general, we can sequentially construct a set of rules to cover instances of ψ, with the new rule as an exception rule to the previous rules.

4.2. Compromising strategies

A compromising strategy promotes the construction of more general rules containing more incorrectly covered exceptions. A compromising strategy means to compromise the accuracy to a certain level, in order to keep another important feature at a relatively high measuring level. That means that the high accuracy is often not the goal in order to preserve or improve another property. Intuitively, a compromising strategy needs to introduce a probability value, say β, to be as an accuracy threshold.

Improve the generality

In most cases, a compromising strategy generates shorter and simpler rules defined by a proper subset of entire feature set. By choosing $A \subseteq At$, a set of formulas ϕ_A can be defined. Borrowing the concept of the rough set theory, we can define a β-positive region with respect to a target concept ψ. The β-positive region is the union of all objects satisfying the rules defined by A with the confidence greater than or equal to β, which is denoted as:

$$POS_A^{\beta}(\psi) \quad = \quad \bigcup \{m(\phi) : P(\psi|\phi) \geq \beta, \phi \in \phi_A\},$$

where $P(\psi|\phi) = \frac{|m(\psi) \cap m(\phi)|}{|m(\phi)|}$. To preserve the generality, a heuristic criterion can be defined as: given a predefined β value, $POS_A^{\beta}(\psi) \geq POS_{At}(\psi)$. A rule set satisfies the generality preservation strategy, with individual rule in form of $\phi \Rightarrow \psi$, can classify more objects in the universe than the set of rules produced by the entire set of At, while keeping the confidence not less than β.

Decrease the cost

A general rule may include more incorrectly covered exceptions. Suppose a set of objects in the universe can be defined by a descriptive formula ϕ. According to a learnt rule, $\phi \Rightarrow \psi_1$. That means that generally all the objects satisfying ϕ should imply ψ_1. However, the rule could be too general, thus, an object $x \in m(\phi)$ implies a decision value different from ψ_1, say, it satisfies a decision value ψ_2. This becomes an exception, or an error, of classification.

For a specific classification exception, denoted as (ψ_1, ψ_2), the exception count is the number of objects in the universe that possess this exception, which can be defined as

$$errCount(\psi_1, \psi_2) = |\{x : des([x]) \Rightarrow \psi_1, des(x) \Rightarrow \psi_2\}|,$$

where $[x]$ indicates the entire equivalence class contains x, $des(.)$ means the description of the given object, or the object set, by a formula, and $\psi_1 \neq \psi_2$.

Yao and Wong applied the Bayesian decision procedure for classification [46]. The basic idea is that different errors may indicate different cost. A rule set satisfying the cost preservation strategy will not increase the error cost.

5. Exploring Explanations of Discovered Pattern

The role of explanation is to clarify, teach, and convince [9]. There are many kinds of explanations. An explanation could be a definition of a term, the cause and effect of an event, or the significance of a phenomenon. Different explanations are the answers to many different kinds of questions. Explanation is both subjective and objective [7]. It is subjective because the meaning of explanation, or the evaluation of a good explanation, is different for different people at different times. On the other hand, explanation is objective because it must win general approval as a valid explanation, or has to be withdrawn in the face of new evidence and criticism. The interpretations and explanations enhance our understanding of the phenomenon and guide us to make rational decisions.

Yao *et al.* suggested to add an explicit explanation module into the existing data mining processes [48]. Explanations of data mining address several important questions, such as, what needs to be explained? How to explain the discovered knowledge? Moreover, is an explanation correct and complete?

5.1. Discovered patterns to be explained

The knowledge discovered from data should be explained and interpreted. Knowledge can be discovered by unsupervised learning methods. Unsupervised learning studies how systems can learn to represent, summarize, and organize the data in a way that reflects the internal structure (namely, a pattern) of the overall collection. This process does not explain the patterns, but describes them. The primary unsupervised techniques include clustering mining, belief networks learning, and association mining. The criteria for choosing which pattern to be explained are directly related to pattern evaluation step of data mining.

5.2. Profiles used to construct explanations

Background knowledge provides features that can possibly explain a discovered pattern. An explanation may include many branches of inquiry: physics, chemistry, meteorology, human culture, logic, psychology, and the methodology of science. In data mining, explanation can be made at a shallow, syntactic level based on statistical information, or at a deep, semantic level based on domain knowledge.

The required information and knowledge for explanation may not necessarily be inside the original dataset. One needs to collect additional information for explanation construction.

The key question is the selection of the features that are generally explanatory to the target concept from many features that happen to be related to the current discovered pattern. Craik [7] argued that the power of explanations involves the power of insight and anticipation. One collects certain features based on the underlying hypothesis that they may provide explanations of the discovered pattern. That something is unexplainable may simply be an expression of the inability to discover an explanation of a desired sort. The process of selecting the relevant and explanatory features may be subjective, and trial-and-error. In general, the better our background knowledge is, the more accurate the inferred explanations are likely to be.

5.3. Explanation construction

Explanations for data mining results reason inductively, namely, drawing an inference from a set of acquired training instances, and justifying or predicting the instances one might observe in the future.

Supervised learning methods can be applied for the explanation construction. The goal of supervised learning is to find a model that will correctly associate the input patterns with the classes. In real world applications, supervised learning models are extremely useful analytic techniques. The widely used supervised learning methods include decision tree learning, rule-based learning, and decision graph learning. The learned results are represented as either a tree, or a set of if-then rules. The constructed explanations give some evidence about under what conditions (within the background knowledge) the discovered pattern is most likely to happen, or how the background knowledge is related to the pattern.

5.4. Explanation evaluation

The role of explanation in data mining is positioned among proper description, relation and causality. Comprehensibility is the key factor in explanations. The accuracy of the constructed explanations relies on the amount of training examples. Explanations perform poorly with insufficient data or poor presuppositions. Different background knowledge may infer different explanations. There is no reason to believe that only one unique explanation exists. One can use statistical measures and domain knowledge to evaluate different explanations.

6. Conclusion

In this paper, we focus on interactive data mining which is characterized by user requirement and user judgement. On the abstract level, we discuss multiple views and user preference in data mining domain. On the application level, we discuss the real problems and concerns while coping with abnormal environments and most-in-need explanations.

We argue that more effective data mining systems should support better human-machine interactivity. The concern of effectiveness and the concern of efficiency should be synchronized with user cognitive phases and requirements. Bearing user requirement in mind, the researches on interactive data mining are fairly broad.

References

[1] Agrawal, R., Imielinski, T. and Swami, A., Mining association rules between sets of items in large databases, *Proceedings of ACM SIGMOD*, 207-216, 1993.

[2] Barber, B. and Hamilton, H., Extracting share frequent itemsets with infrequent subsets, *Data Mining and Knowledge Discovery*, 7, 153-185, 2003.

[3] Brachmann, R. and Anand, T., The process of knowledge discovery in databases: a huamn-centered approach, *Advances in Knowledge Discovery and Data Mining*, AAAI Press & MIT Press, Menlo Park, CA, 37-57, 1996.

[4] Brin, S., Motwani, R. and Silverstein, C., Beyond market baskets: generalizing association rules to correlations, *Proceedings of ACM SIGMOD*, 265-276, 1997.

[5] Chen, Y.H. and Yao, Y.Y. Multiview intelligent data analysis based on granular computing, *Proceedings of IEEE International Conference on Granular Computing (Grc06)*, 281-286, 2006.

[6] Compton, P. and Jansen, B., Knowledge in context: a strategy for expert system maintenance, *Proceedings of the 2nd*

Australian Joint Conference of Artificial Intelligence, 292-306, 1988.

[7] Craik, K., *The Nature of Explanation*, Cambridge University Press, London, New York, 1943.

[8] Demri, S. and Orlowska, E., Logical analysis of indiscernibility, in: E. Orlowska, (Ed.), *Incomplete Information: Rough Set Analysis*, Physica-Verlag, Heidelberg, 347-380, 1998.

[9] Dhaliwal, J.S. and Benbasat, I., The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation, *Information Systems Research*, 7, 342-362, 1996.

[10] Doyle, J., Rationality and its role in reasoning, *Computational Intelligence*, 8, 376-409, 1992.

[11] Elm, W.C., Cook, M.J., Greitzer, F.L., Hoffman, R.R., Moon, B. and Hutchins, S.G., Designing support for intelligence analysis, *Proceedings of the Human Factors and Ergonomics Society*, 20-24, 2004.

[12] Fishburn, P.C., *Utility Theory for Decision-Making*, John Wiley & Sons, New York, 1970.

[13] Gago, P. and Bento, C., A metric for selection of the most promising rules, *Proceedings of PKDD*, 19-27, 1998.

[14] Ganter, B. and Wille, R., *Formal Concept Analysis: Mathematical Foundations*, Springer-Verlag, New York, 1999.

[15] Han, J., Hu, X. and Cercone, N., A visualization model of interactive knowledge discovery systems and its implementations, *Information Visualization*, 2, 105-125, 2003.

[16] Hancock, P.A. and Scallen, S.F., The future of function allocation, *Ergonomics in Design*, 4, 24-29, 1996.

[17] Hilderman, R.J. and Hamilton, H.J., *Knowledge Discovery and Measures of Interest*, Kluwer Academic Publishers, Boston, 2001.

[18] Jung, S.Y., Hong, J.H. and Kim, T.S., A formal model for user preference, *Proceedings of ICDM'02*, 235-242, 2002.

[19] Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A., *Foundations of Measurement*, Academic Press, New York, 1971.

[20] Lee, T.T., An information-theoretic analysis of relational databases - part I: data dependencies and information metric, *IEEE Transactions on Software Engineering*, 13, 1049-1061, 1987.

[21] Liu, B., Hsu, W. and Ma, Y., Pruning and summarizing the discovered associations, *Proceedings of KDD*, 125-134, 1999.

[22] Liu, B., Hsu, W. and Chen, S., Using general impressions to analyze discovered classification rules, *Proceedings of KDD*, 31-36, 1997.

[23] Mari, D.D. and Kotz, S., *Correlation and Dependence*, Imperial College Press, London, 2001.

[24] Pawlak, Z., Rough sets, *International Journal of Computer Information and Science*, 1982, 11(5), 341-356.

[25] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Dordrecht, 1991.

[26] Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules, in: G. Piatetsky-Shapiro and W.J. Frawley (Eds.), *Knowledge Discovery in Databases*, AAAI/MIT Press, 229-238, 1991.

[27] Ras, Z. and Wieczorkowska, A., Action rules: how to increase profit of a company, *Proceedings of PKDD*, 587-592, 2000.

[28] Robers, F., *Measurement Theory*, Addison Wesley, Massachusetts, 1979.

[29] Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, third edition, Addison-Wesley, 1998.

[30] Silberschatz, A. and Tuzhilin, A., On subjective measures of interestingness in knowledge discovery, *Proceedings of KDD*, 275-281, 1995.

[31] Silberschatz, A. and Tuzhilin, A., What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering*, 8, 970-974, 1996.

[32] Smyth, P. and Goodman, R., An information theoretic approach to rule induction from databases, *IEEE Transactions on Knowledge and Data Engineering*, 4, 301–316, 1992.

[33] Tan, P.N., Kumar, V. and Srivastava, J., Selecting the right interestingness measure for association patterns, *Proceedings of KDD*, 2002.

[34] Wang, K. and He, Y., User-defined association mining, *Proceedings of PAKDD*, 387-399, 2001.

[35] Wang, K., Zhou, S. and Han, J., Profit mining: from patterns to actions, *Proceedings of EDBT*, 70-87, 2002.

[36] Wang, Y.X., On cognitive informatics, *Proceedings of ICCI'02*, 34-42, 2002.

[37] Wang, Y.X. and Liu, D., On information and knowledge representation in the brain, *Proceedings of ICCI'03*, 26-29, 2003.

[38] Wang, Y.X., On autonomous computing and cognitive processes, *Proceedings of ICCI'04*, 3-4, 2004.

[39] Wille, R., Restructuring lattice theory: an approach based on hierarchies of concepts, in: I. Rival (Ed.), *Ordered sets*, Reidel, Dordecht-Boston, 445-470, 1982.

[40] Wille, R., Concept lattices and conceptual knowledge systems, *Computers Mathematics with Applications*, 23, 493-515, 1992.

[41] Yao, Y.Y., On modeling data mining with granular computing. *Proceedings of COMPSAC*, 638-643, 2001.

[42] Yao, Y.Y., Information-theoretic measures for knowledge discovery and data mining, in: Karmeshu (Ed.), *Entropy Measures, Maximum Entropy and Emerging Applications*, Springer, Berlin, 115-136, 2003.

[43] Yao, Y.Y., Mining high order decision rules, in: M. Inuiguchi, S. Hirano and S. Tsumoto (Eds.), *Rough Set Theory and Granular Computing*, Springer, Berlin, 125-135, 2003.

[44] Yao, Y.Y., Perspectives of granular computing, *Proceedings of 2005 IEEE International Conference on Granular Computing*, 1, 85-90, 2005.

[45] Yao, Y.Y., Chen, Y.H. and Yang, X.D., A measurement-theoretic foundation for rule interestingness evaluation, *Proceedings of Workshop on Foundations and New Directions in Data Mining in the 3rd IEEE International Conference on Data Mining (ICDM 2003)*, 221-227, 2003.

[46] Yao, Y.Y. and Wong, S.K.M., A decision theoretic framework for approximating concepts, *International Journal of*

Man-machine Studies, 37, 793-809, 1992.

[47] Yao, Y.Y., Wang, F.Y., Wang, J. and Zeng, D., Rule + exception strategies for security information analysis, *IEEE Intelligent Systems*, 20, 52-57, 2005.

[48] Yao, Y.Y., Zhao, Y. and Maguire, R.B., Explanation-oriented association mining using rough set theory. *Proceedings of Rough Sets, Fuzzy Sets and Granular Computing*, 165-172, 2003.

[49] Yao, Y.Y. and Zhong, N., An analysis of quantitative measures associated with rules, *Proceedings of PAKDD*, 479-488, 1999.

[50] Zhao, Y. and Yao, Y.Y., Interactive user-driven classification using a granule network, *Proceedings of ICCI'05*, 250-259, 2005.

[51] Zhong, N., Yao, Y.Y. and Ohshima, M., Peculiarity oriented multi-database mining, *IEEE Transactions on Knowlegde and Data Engineering*, 15, 952-960, 2003.

[52] Zhong, N., Yao, Y.Y., Ohshima, M. and Ohsuga, S., Interestingness, peculiarity, and multi-database mining, *Proceedings of IEEE International Conference on Data Mining (ICDM'01)*, 566-573, 2001.