World Academy of Science, Engineering and Technology
International Journal of Industrial and Manufacturing Engineering
Vol:1, No:11, 2007

Development of Subjective Measures of Interestingness: From Unexpectedness to Shocking

Eiad Yafi, M. A. Alam, and Ranjit Biswas

***Abstract*—**Knowledge Discovery of Databases (KDD) is the process of extracting previously unknown but useful and significant information from large massive volume of databases. Data Mining is a stage in the entire process of KDD which applies an algorithm to extract interesting patterns. Usually, such algorithms generate huge volume of patterns. These patterns have to be evaluated by using interestingness measures to reflect the user requirements. Interestingness is defined in different ways, (i) Objective measures (ii) Subjective measures. Objective measures such as support and confidence extract meaningful patterns based on the structure of the patterns, while subjective measures such as unexpectedness and novelty reflect the user perspective. In this report, we try to brief the more widely spread and successful subjective measures and propose a new subjective measure of interestingness, i.e. shocking.

***Keywords*—**Shocking rules (SHR).

I. Introduction

COLLECTING the data for mining is very hard process by itself as the ongoing operations everyday generate tremendous and huge amount of data. Data Mining helps the end users extract interesting business information or patterns from large databases, and the larger the volume of data that can be processed by data mining techniques, the greater the confidence in the result [1,3]. Data mining process can be of one or more of the following functions such as classification rules, regression, time series analysis, prediction, clustering, summarization, association rules and sequence discovery. The number of generating rules would be very high and only few of the discovered patterns are of the interest to the end user. Many researchers have identified some measures of interestingness of discovered rules. These measures are support, confidence, statistical significance, simplicity [6] and these measures are called objective measures and unexpectedness, actionability and novelty which are called subjective measures.

In this survey, we will list all the subjective measures of interestingness which have been introduced before and we will add another subjective measure to the previously proposed measures**.**

Eiad Yafi is a research scholar from Syria, working on his PhD in Data Mining at Hamdard University, New Delhi, Iran (e-mail: eiad.yafi@ gmail.com).

M. A. Alam is a professor at Hamdard University, New Delhi, India (e-mail: alam@jamiahamdard.ac.in).

Ranjit Biswas is a visiting professor at Hamdard University, New Delhi, India (e-mail: ranjitbiswas@yahoo.com).

II. Subjective Measures

Objective measures of interestingness may not highlight the most important patterns produced by the data mining system, subjective measures generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. It should be noted that both objective and subjective measures should be used to select interesting rules. Objective measures can be used as a kind of first filter while subjective measures can be used as a final filter to elect truly interesting rules [11]. Identifying interesting rules from a set of discovered rules is not a simple task because a rule could be interesting to one user but of no interesting to another. The interestingness of a rule is a subjective matter because it depends on the user's existing concepts and information about the domain and user's interest. Three main subjective measures of interestingness are *unexpectedness* (Silberschatz and Tuzhilin 1996) and (Bing Liu 1997) and *actionability* (Piatesky-Shapiro and Matheus 1994a) and *novelty* (Silberschatz and Tuzhilin and Freitas).

A. Unexpectedness

[3] Studied the interestingness of discovered rules in a healthcare domain. The author in [4] introduced KAFIR and argued that a good measure of the interestingness of a finding is the estimated benefit that could be realized by taking a specific action in response. KAFIR classifies all possible findings into a predefined set of types, and then it defines the production rule of finding that specifies the actions to be taken. The analyst provides recommendation based on his prior knowledge. The system determines all the production rules matching this finding and selects the rule with the highest probability of success. However, this system is considered to be good but can't be used for any other application as it's a domaindependent. Unexpectedness has been studied in [4] as a probabilistic measure. Silberschatz and Tuzhilin devised a system of *hard* and *soft* beliefs. The belief is used for defining unexpectedness. A rule is considered interesting if it surprises the user and contradict an exist belief. The soft beliefs could be revised and updated in the light of new evidence but the hard beliefs are constraints and cannot be changed with new evidence, and if new evidence contradicts hard beliefs it

would tend to some mistakes made in acquiring this new evidence.

In [8] a new approach was presented based on syntactic distance technique. It measures the distance between the new rule and a set of beliefs. The rule and the belief are dissimilar if either the consequents are same but antecedents are far apart or antecedents are same but consequent are far apart. The rule is unexpected in case of either unexpected consequent or unexpected condition.

Another definition of unexpectedness was presented in [9, 10] based on logical contradiction. This technique was devised for association rules which use the statistical strength (support) of a rule to determine if a rule is unexpected or not. A rule $A \to B$ is unexpected with respect to the belief $X \to Y$ if the heads A and B contradict each other logically and the rule $A, X \to B$ holds.

Also in [7], a general impression is used to evaluate the importance of classification rules by comparing the discovered rules against the user's existing concepts or vague feelings. The user specifies all the general impressions that he/she has about the domain and the system analyzes the discovered rules by matching them against the impressions using fuzzy matching approach. The discovered rules are then ranked and unexpectedness is measured.

B. Actionability

Actionability is an important subjective measure of interestingness because users are mostly interested in the knowledge that permits them to do their jobs better by taking some specific actions in response to the newly discovered knowledge [11]. Silberschatz and A. Tuzhilin were first to discuss actionability in 1996, they stated that a rule is interesting (actionable) if the user can do an action to his/her advantage. It's well known that most of data mining algorithms deal with business activities. Action rules are all about the profit and to improve the business policies. Profit mining approach was presented in [13]; the key to profit mining is to recommend right item and right prices. If the price is too high, the customer will go away without generating any profit, if the price is too low or if the item is not profitable, the profit will not be maximized.

C. Novelty

What makes the KDD process successful is whether it extracts previously, unknown, useful and interesting knowledge [4, 12]. No much work has been done on novelty as subjective measures it was studied in some other disciplines such as robotics, machine learning and statistical outlier detection. Novelty was studied in [14] for detecting novel rules mined from text. The novelty is estimated based on the Lexical knowledge in WordNet. The proposed approach defines a measure of semantic distance between two words. The novelty is defined as the average distance across all pairs of words (w_i, w_j) where w_i a word in the antecedent is and w_j is a word in the consequent.

In [15] Alhegami and Bhatnagaar proposed a framework to quantify the novelty in terms of the computing the deviation of currently discovered knowledge with respect to the previously discovered knowledge. A rule is novel if, to some extent contributes to new knowledge. They proposed a technique to measure the deviation at conjunct level and then it's generalized to rule level. A rule is comparable with another if they have the same consequent otherwise the discovered rule is having the highest degree of deviation. A rule deviation is calculated as a linear combination of deviation of the set of conjuncts of the rule. The novelty of a rule then depends on the user feeling about the domain which is determined by a certain threshold value.

III. SHOCKING RULES

After listing all the available subjective measures of interestingness we introduce another measure of rule interestingness that is shocking rules. The idea of shocking rules came from the latest disasters which have encountered the world recently such as the increasing number of earthquakes, tornados and Tsunami waves. The more interesting rules are those which are unexpected and novel as well, so shocking rules have the highest degree of interestingness, shocking rules are novel since they do not exist in the previously discovered knowledge (PDK) and are at the same considered unexpected as they have the highest deviation from some rules in PDK. A rule $R: A \to B$ is shocking if it overthrows all the expectations of the user. It's unprecedented, never expected and happens suddenly in a way that it shocks the user and put him in an unenviable situation. What make the shocking rules and surprising rules different is the massive decrease /increase in the value of a conjunct. Let's take the example of Tsunami, consider the conjunct $X = V$ where X implies the attribute of the height of a sea wave and V is the value. Now every time this conjunct is being matched with a set of beliefs in PDK this conjunct is expected if V is in the range {a to b}, where a is the minimum value ever recorded and b is the highest. This conjunct is unexpected or surprising if a slight change in v happens. But when an unprecedented change in v takes place then we call this conjunct a shocking conjunct.

As we stated before this rule does not exist in PDK and we can't update the PDK with such rule so we suggest adding such rules to new domain knowledge so that we can calculate the degree of shocking rules.

A. Definitions

A rule R has the form: $A \to C$ where A denotes as antecedent and C denotes a consequent. We consider both A and C as sets of conjuncts in our research work.

Shocking rules can be defined in two major cases:

- If $\Psi(A_1, A_2) = \varepsilon \to \Psi(C_1, C_2) \geq \alpha$ then $c_1 \in A_2$ is Significant
- If $\Psi(A_1, A_2) \geq \alpha \to \Psi(C_1, C_2) = \varepsilon$ then $c_1 \in A_2$ is Significant,

where α is a threshold value determined by the user, A_1, C_1 are set of conjuncts from a rule R_1 represents it's antecedent and consequent respectively and A_2, C_2 are set of conjuncts from a rule R_2 represents it's antecedent and consequent respectively.

.

Significance of attribute can be defined as:

Significance (Attribute) = $\begin{Bmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix}$ where these values indicate the significance's degree of an attribute as following table:

TABLE I
SIGNIFICANCE OF ATTRIBUTE

Value	Indication	Comments
0	Normal	No change in the class value
x_1	Low Significance	Change in the attribute value
x_2	Intermediate Significance	Change in the attribute operator
x_3	High Significance	Change in the attribute operator and value
x_4	Very High	Change in the attribute name

IV. EXPERIMENT

In this work, we are trying to measure the degree of significance of significant attributes and then the degree of shocking rules (SHR) from a natural disaster dataset, but we were unable to get such a real dataset, so we are trying to build up an artificial dataset and assume the type of shocking rules we are supposed to get as no data mining classifier is able to extract such rules.

The importance of this work lies on the need to build a warning system for natural disasters such as Tsunami and reduce the time needed to save lives and reduces the risks.

We hope the implementation of our artificial dataset will be ready soon to be the first to introduce such measure to be used in real life applications.

REFERENCES

[1] Alex Berson & Stephen Smith, Data Warehousing & Data Mining, Hill Edition 2004.Ding, W., and Marchionini, G. *A Study on Video Browsing Strategies.* Technical Report UMIACS-TR-97-40, University of Maryland, College Park, MD, 1997.

[2] Harry Singh, Data Warehousing, concepts, technologies, implementations, 1998.

[3] G. Piatesky-Shapiro and C. J Matheus, The interestingness of deviations. In Proceedings of AAAI.

[4] A. Silberschatz and A. Tuzhilin, On subjective measures of interestingness in knowledge discovery., Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995, 275-281.

[5] B. Liu & W. Hsu, L. Mun, and H. lee, Identifying Interesting Missing patterns.

[6] B. Liu & W. Hsu, L. Mun, and H. Lee, Finding interesting patterns using user expectations, *Technical Report TRA7/*96, Department of Information Systems and Computer Science, National University of Singapore, 1996.

[7] B. Liu, W. Hsu, and S. Chen, Using general impressions to analyze discovered classification rules.

[8] Bing Liu & Wynne Hsu, Post Analysis of Learned Rules, *AAAI-*96, Aug 4-8, 1996, Portland, Oregon, USA.

[9] B. Padmanabhan and A. Tuzhilin, Unexpectedness as a measure of interestingness in knowledge discovery.

[10] B. Padmanabhan and A. Tuzhilin, A belief-driven method for discovering unexpected patterns.

[11] Zengyou He, Xiaofei Xu and Shengchun Deng, Data mining for actionable knowledge: a survey.

[12] A. Silberschatz and A. Tuzhilin, what makes patterns interesting in knowledge discovery systems.

[13] Ke Wang, Senqiang Zhou and Jiawei Han, Profit mining: from patterns to actions.

[14] Basu, S., Mooney, R.,J .,Pasupuleti,K. V.,Ghosh,J.: using Lexical knowledge to evaluate the novelty of rules mined from text. In proceedings of the NAACL workshop and other Lexical resources: Applications, Extentions and customization

[15] A. S. Al-Hegami, V. Bhatnagar, N. KumarNovelty Framework for Knowledge Discovery in Databases, DaWaK 2004: 48-57.

[16] F. Hussain, H. Liu, E. Suzuki and H. Lu, Exception rule mining with a relative interestingness measure.

[17] E. Suzuki, Discovering unexpected exceptions: A stochastic approach. In Proc. RFID, pages 225-232, 1996

[18] E. Suzuki, Discovery of surprising exception rules based on intensity on implication. In proc. Second Pacific-Asia conference on Knowledge Discovery and data mining (PAKDD), 1998.

[19] R. Agarwal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases. In Proc. Of the ACM SIGMOD Conference on Management of data, Washington, D.C., May 1993.

.