International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

A Hybrid Approach for Quantification of Novelty in Rule Discovery

Vasudha Bhatnagar, Ahmed Sultan Al-Hegami, and Naveen Kumar

Abstract— Rule Discovery is an important technique for mining knowledge from large databases. Use of objective measures for discovering interesting rules lead to another data mining problem, although of reduced complexity. Data mining researchers have studied subjective measures of interestingness to reduce the volume of discovered rules to ultimately improve the overall efficiency of KDD process.

In this paper we study *novelty* of the discovered rules as a subjective measure of interestingness. We propose a hybrid approach that uses objective and subjective measures to quantify *novelty* of the discovered rules in terms of their deviations from the known rules. We analyze the types of deviation that can arise between two rules and categorize the discovered rules according to the user specified threshold. We implement the proposed framework and experiment with some public datasets. The experimental results are quite promising.

***Keywords*—** Knowledge Discovery in Databases (KDD), Data Mining, Rule Discovery, Interestingness, Subjective Measures, Novelty Measure.

I. INTRODUCTION

The vast search space of hidden patterns in the massive databases is a challenge for the KDD community. For example, in a database with n distinct items, the number of potential frequent item sets is exponential in n. In a database with n records, the potential number of clusters is $\frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}(i)^n$ [6]. However, a vast majority of these patterns are pruned by the score functions engaged in the mining algorithm. To avoid computing the score function for the entire search space, optimization strategies are used. For example, in association rule mining, confidence is the commonly used score function and the anti monotonic property of frequent itemsets is the optimization strategy.

Despite massive reduction of search space by employing suitable score function and optimization strategies, all of the discovered patterns are not useful for the users. Consequently, researchers have been strongly motivated to further restrict the search space, by putting constraints [5] and providing good measures of interestingness [9,18].

Constraints based mining techniques allow the users to specify the rules to be discovered according to their background knowledge, thereby making the KDD process more effective [5,10]. A complicated mining query can be used to express these constraints specified by the user in order to make the mining process more efficient.

There are two aspects of interestingness measures that have been studied in data mining literature viz. Objective and Subjective measures. Objective measures are based on the statistical significance (certainty, coverage, etc.) or structure (simplicity) of the patterns [7,10]. Subjective measures are based on the user who evaluates the patterns such as novelty, actionability and unexpectedness, etc. [4,9,12,3,13,2].

In real life KDD endeavors, it is often required to compare the rules mined from datasets generated under different context (for example, at different points in time or in two different locations). Unless the underlying data generation process has changed dramatically, it is expected that the rules discovered from one set are likely to be similar (in varying degrees) to those discovered from another set. Some of the discovered rules may be identical to the known rules, some may be generalization/ specialization of the known rules and some others may be same or different with varying degrees of sameness/difference.

As the numbers of rules discovered by data mining algorithms become huge, the time consumed and the space required for maintaining and understanding these rules becomes vast. *Novelty* of a rule can be used as an effective way of reducing the size of the rule set discovered from the target data set.

Novelty is defined as the extent to which the discovered rules contribute to new knowledge [1,2,3]. In this paper we focus on the quantification of *novelty* and use this measure for categorization of discovered rules. Though *novelty* is a subjective measure, we propose a strategy to quantify objectively the *novelty index* of each discovered rule, and facilitate categorization of rules with degree of novelty desired by the user. Asking the user to specify a threshold to filter rules of desired degree of novelty captures user subjectivity.

II. RELATED WORK

There are many proposals that studied the novelty in disciplines such as robotics, machine learning and statistical outliers detection [14,15,16,17]. Generally, these methods

Vasudha Bhatnagar is lecturer in the Department of Computer Science, University of Delhi, Delhi, INDIA; (e-mail: vb@csdu.org).
Ahmed Sultan Al-Hegami is Ph.D scholar in the same department; mobile: +91-9811122973; (e-mail: ahmed_s_gamil@yahoo.com)
Naveen Kumar is reader in the same department ; (e-mail: nk@csdu.org).

build a model of training set that is selected to contain no examples of the important (i.e. novel) class [11].

To best of our knowledge, no concrete work has been conducted to tackle the novelty measure of rules discovered by data mining algorithms. The work that has been proposed is detecting the *novelty* of rules mined from text [8]. *Novelty* is estimated based on the lexical knowledge in WordNet [8]. The proposed approach defines a measure of semantic distance between two words in WordNet and determined by the length of the shortest path between the two words (w_i,w_j). The novelty then is defined as the average of this distance across all pairs of the words (w_i,w_j), where w_i is a word in the antecedent and w_j is a word in the consequent.

In [2], we proposed a framework to quantify the novelty in terms of computing the deviation of currently discovered knowledge with respect to domain knowledge and previously discovered knowledge. The approach presented in [2] is intuitive in nature and lays more emphasis on user involvement in quantification process by way of parameter specification. In the present work, the quantification is performed objectively and user involvement is sought in categorization of rules based on *novelty index*.

III. Novelty Index

Let D_i denote the database extension at time t_i, and k_i denote the knowledge discovered. Then, the shaded portions of Figure 1 denote the knowledge carrying high degree of novelty. Major volume of k_i would be the overlapping region that represents previously discovered knowledge. Thus the rules falling in the shaded area are assigned high degree of *novelty* compared to those in the overlapping regions.

The proposed framework assigns a *novelty index* to each discovered rule that indicates its proximity/deviation from some existing rule in the rule base of previously discovered knowledge.

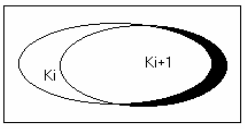

Figure 1. Regions of Discovered Rules

Novelty index of a rule is the deviation with respect to a given rule set. It is a pair (Ă, Č) that indicates the deviation of the antecedent and consequent of the rule with those of the closest rule in the previously discovered knowledge. To compute the *novelty index* of a rule, we measure the deviation for the antecedent and the consequent at conjunct level and subsequently, combine the conjunct level deviation to compute rule level deviation.

A. Definitions and Notations

A rule *R* has the form: A→C where A denotes an antecedent and C denotes a consequent. Both A and C are in CNF ($c_1 \Lambda c_2 \Lambda \ldots\ldots \Lambda c_k$). The conjunct c_j is of the {Å, O, V}. Where Å is an attribute, *Dom* (Å) is the domain of Å, and V ∈ *Dom* (Å), O ∈ {=,<,>,≥,≤}. Without loss of generality, we consider both A and C as sets of conjuncts for further processing.

B. Deviation at Conjunct Level

In order to quantify deviation between any two conjuncts, the attributes, operators, and attribute values of the two conjuncts in question need to be taken into account.

Definition 3.1 Two conjuncts c_i and c_j ($Å_iO_iV_i$ and $Å_jO_jV_j$ respectively) are compatible if and only if $Å_i = Å_j$. Otherwise, we consider c_i and c_j as non-compatibles.

Definition 3.2 Let c_i and c_j be two non-compatible conjuncts. The deviation δ (c_i,c_j) between them is defined to be 1.

We capture the following four types of deviations between two compatible conjuncts.

Z-deviation: This type signifies identical conjuncts and is quantified by numeric 0.

V-deviation: This type of deviation signifies the magnitude of change in the value of the attribute in two conjuncts. In order to normalize, we quantify this type of deviation as the ratio of the change to the range of the attribute value.

This method of computation of V-deviation is suitable for only numeric and ordinal attributes. In case of nominal attributes, the change in value can be quantified in terms of probabilities. Since ordinal domains generally have small and manageable cardinality, prior domain knowledge can be used to assign probabilities to domain values. In case it is not feasible to assign probabilities in the above-mentioned way (e.g. color of car), the dataset itself can be used to compute probabilities corresponding to each domain value.

C-deviation: This type of deviation signifies the deviation in the conditional operators in the two conjuncts. In order to quantify C-deviation, we take into account the type of change in the condition. The operators are formatted on a number line as shown in Figure 2. The deviation between the operators is quantified by the distance between the operators on the numberline.

We define a function *opdist* (O_1, O_2) → {1,2,3,4}, which denotes the distance between the two distinct operators (O_1, O_2) on the numberline. We define four possible values of deviations (1/5, 2/5, 3/5, 4/5) between any two operators, ranking the extent of deviation between condition operators in two conjuncts.

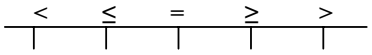

Figure 2. Operators on Numberline

CV-deviation: Quantifies V-deviation in presence of C-deviation. It captures the co-occurrence of change in both conditions and attribute values in two conjuncts.

We compute the C-deviation (*c*) and V-deviation (*v*) independently of each other, in the two given conjuncts. The user defines a real valued function $f(c,v) \rightarrow [0,1]$ to combine the two types of deviations. Depending on the importance of the type of deviations for a specific application in a domain, different functions can be used for computing deviations on different attributes.

Typically, $f(c,v)$ is of the form w_1c+w_2v, where

$w_1+w_2=1$.

Note that, the computation of deviation between two conjuncts is objective in all types of deviation, except CV-deviation, where the user subjectivity is captured. Depending on the importance of either C or V deviation, the user assigns appropriate weights w_1 and w_2.

The following definition formalizes the quantification of Conjunct level deviation.

Definition 3.3 Let c_1 and c_2 be two compatible conjuncts ($A_1O_1V_1$ and $A_2O_2V_2$ respectively). The deviation of c_1 with respect to c_2 is defined as follows:

$$\delta = (c_1, c_2) = \begin{cases} 0 & \text{if } O_1 = O_2 \text{ and } V_1 = V_2 \quad Z\text{ - deviation.} \\ \dfrac{|V_1 - V_2|}{Range(Dom(A_1))} & \text{if } O_1 = O_2 \text{ and } V_1 \neq V_2 \quad V\text{ - deviation.} \\ opdist(O_1, O_2)/5 & \text{if } O_1 \neq O_2 \text{ and } V_1 = V_2 \quad C\text{ - deviation.} \\ f(c,v) & \text{if } O_1 \neq O_2 \text{ and } V_1 \neq V_2 \quad CV\text{ - deviation.} \end{cases}$$

Lemma 3.1. The conjunct level deviation lies between [0,1]

Proof 3.1. By definitions 3.2 and 3.3.

It is easy to see that: i) $\delta(c_i,c_j) \geq 0$, ii) $\delta(c_i,c_i)=0$,

iii) $\delta(c_i,c_j)= \delta(c_j,c_i)$. However, $\delta(c_i,c_j)$ does not satisfy triangular inequality in case of on CV-deviation, where we capture user subjectivity

C. Conjunct Set Deviation

In order to quantify the deviation $\Psi(S_1,S_2)$ between two conjuncts sets, we analyze the possible types of differences between two sets of conjuncts S_1 and S_2. Without loss of generalization, we assume that an attribute occurs at most once in a conjunct set S. Computation of deviation at this level is based on counting incompatible conjunct among the two sets and quantifying total deviation among the compatible conjuncts. Intuitively, it is the number of incompatible conjuncts that contribute most towards the value of the deviation. We compute the deviation between two conjunct sets as follows.

Definition 3.5 Let S_1 and S_2 be two conjunct sets with cardinalities $|S_1|$ and $|S_2|$ respectively. Let k be the pairs of compatible conjuncts between S_1 and S_2. The deviation between S_1 and S_2 is computed as:

$$\Psi(S_1,S_2)= \frac{\{|S_1|+|S_2|-2*k\}+\sum_{i=1}^{k}\delta(c_1^i,c_2^i)}{|S_1|+|S_2|}$$

where (c_1^i, c_2^i) is the i^{th} pair of compatible conjuncts.

Theorem 3.1 For any two conjunct sets S_1 and S_2, $0 \leq \Psi(S_1,S_2) \leq 1$.

Proof 3.1 The proof follows by simple reasoning. We consider two extreme cases where there are no compatible conjuncts and another with all equal conjuncts.

In case there are no compatible conjuncts, $k=0$ and the second component of the numerator vanishes. With all non-compatible conjuncts $\Psi(S_1,S_2)=1$. In case the two conjunct sets are equal, $k = \frac{|S1|+|S_2|}{2}$ and the second component in the numerator reduces to zero. Thus $\Psi(S_1,S_2) = 0$, which captures Z-deviation.

Note that, i)$\Psi(S_1,S_2)\geq 0$,ii)$\Psi(S_1,S_1)= 0$, iii)$\Psi(S_1,S_2)= \Psi(S_2,S_1)$.

We do not expect Ψ to satisfy triangular inequality in view of its violation by underlying conjunct level deviation function. Therefore, Ψ can't be used as a distance metric.

IV. Computing Novelty Index

Novelty index of a rule r is defined with respect to a given rule set R. It is computed as paired deviation of antecedent and consequent of r relative to the closest rule in R. The rule $s \in R$, from whose antecedent the deviation of r is minimum is considered to be closest. The novelty index is defined as follows.

Definition 3.5 Let r: $A_r \rightarrow C_r$ be a rule whose *novelty index* is to be computed with respect to the rule set R. Then

$N_r^R = (\min_{s \in R} (\Psi(A_r,A_s)), \Psi(C_r,C_s))$ gives the *novelty index.*

Having computed the *novelty index* for all the rules in the currently discovered rule set with respect to previously discovered rule set, the task of rule reduction can be performed in several ways. Some of the suggested ways are:
i) select the top K novel rules, ii) select rules with novelty index exceeding a threshold, iii) categorize the indexed rules as per Table I ($\check{A} = \Psi(A_r,A_s)$ & $\check{C} = \Psi(C_r,C_s)$ & Φ is user specified threshold).

TABLE I
CATEGORIZATION OF DISCOVERED RULES

Categorization	Semantics	Condition
Conforming Rules	Rules that have been discovered earlier.	$\check{A} \leq \Phi$ & $\check{C} \leq \Phi$
Generalized (Specialized) Rules	Rules that are generalization (specialization) of some earlier discovered rules.	$A_r(A_s)$ subsumes $A_s(A_r)$ & $\check{C}=0$
Unexpected Rules	Rules that are unexpectedly different from the previously discovered rules.	$\check{A} \leq \Phi$ & $\check{C} \geq \Phi$ OR $\check{A} \geq \Phi$ & $\check{C} \leq \Phi$
Novel Rules	Rules that add to knowledge. Such rules do not fall into any of the earlier specified categories.	$\check{A} \geq \Phi$ & $\check{C} \geq \Phi$

V. Experimental Study

Our proposed approach is implemented in C language and tested using public datasets [19]. Since, there are no other approaches available, which handle the novelty; we could not perform any comparison against our approach. We will show the effectiveness of our framework through the following experiments.

A. Experiment 1

We worked with five public datasets available at [19]. We considered each of these datasets as evolving with time, and partitioned them into 3 increments; D_1, D_2 and D_3 mined at times T_1, T_2 and T_3 respectively. We took each of these partitions to be equal for purpose of generating rules.

The datasets were mined using CBA [20], with 0.1% and 1% as minimum confidence and support respectively, uniformly for all datasets. The low thresholds enable generation of large number of rules; thereby demonstrating the effectiveness of the framework. The discovered rules we categorized as in Table I, with Φ=0.5 and $\mathcal{f}(c,v) \rightarrow$ [0.4,0.6] for CV-deviation. The result is summarized in Table II.

TABLE II
DISCOVERED RULES AT TIME T_1, T_2 AND T_3 FOR DIFFERENT DATASETS WITH Φ=0.5

Dataset	Time	Instances	Discovered rules	Novel	Unexpected	Specialized	Generalized	Conformed
Census	T_1	12000	942	29	239	4	19	652
	T_2	12000	1061	6	189	20	21	825
	T_3	8561	636	3	58	8	7	560
Supmart	T_1	40	2775	25	1576	62	75	1025
	T_2	40	1875	0	1026	49	103	661
	T_3	48	1570	0	717	40	116	697
German	T_1	333	117	13	66	0	0	38
	T_2	333	118	9	43	0	0	66
	T_3	334	133	4	56	3	1	69
Sick	T_1	933	29	4	18	0	1	6
	T_2	933	33	2	17	7	0	7
	T_3	934	32	2	16	5	1	8
Heart	T_1	90	38	7	5	24	0	2
	T_2	90	95	6	6	71	2	10
	T_3	90	40	2	4	19	2	13

TABLE III
DISCOVERED RULES AT TIME T_1, T_2 AND T_3 FOR DIFFERENT (Φ)

Novelty Degree (Φ)	Time	Discovered rules	Novel	Unexpected	Specialized	Generalized	Conformed
Φ=0.9	T_1	942	4	318	0	2	618
	T_2	1061	0	451	5	2	603
	T_3	636	0	241	6	1	388
Φ=0.8	T_1	942	6	241	1	4	690
	T_2	1061	1	235	1	2	822
	T_3	636	0	130	4	1	501
Φ=0.7	T_1	942	10	325	1	5	601
	T_2	1061	2	314	4	6	735
	T_3	636	0	164	5	2	465
Φ=0.6	T_1	942	16	227	15	11	673
	T_2	1061	7	135	16	16	887
	T_3	636	1	79	16	9	531
Φ=0.5	T_1	942	29	239	4	19	652
	T_2	1061	6	189	20	21	825
	T_3	636	3	58	8	7	560

B. Experiment 2

The second experiment was performed using 'census' dataset to study the effect of novelty threshold Φ on the number of rules of different categories. This dataset contains 48842 instances, mix of continuous and discrete attributes, and 2 class values. With same partitions (12000,12000,8561) and support and confidence thresholds as in the previous experiment. The number of rules varied as per our expectation. The result is shown in Table III.

VI. Conclusion

In this paper, we proposed a strategy for rule set reduction based on the *Novelty index* of the rule. *Novelty index* of a newly discovered rule is the quantification of its deviation with respect to the known rule set. User subjectivity is captured by specification of threshold(s) for rule categorization.

The framework is implemented and evaluated using real-life datasets and results have been presented. The generated rules were categorized as conforming, generalized/specialized, unexpected and novel rules.

.

REFERENCES

[1] A. S. Al-Hegami, " Subjective Measures and their Role in Data Mining Process ", In Proceedings of the 6th International Conference on Cognitive Systems, New Delhi, India, 2004.
[2] A. S. Al-Hegami, V. Bhatnagar, and N. Kumar, " Novelty Framework for Knowledge Discovery in Databases ", In Proceedings of 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2004), Spain, 2004.
[3] A. S. Al-Hegami, " Interestingness Measures of KDD : A Comparative Analysis ", In Proceedings of the 11th International Conference on Concurrent Engineering: Research and Applications, China, 2004.
[4] B. Padmanabhan and A. Tuzhilin, " Unexpectedness as a Measure of Interestingness in Knowledge Discovery ", Working paper # IS-97-6, Dept. of Information Systems, Stern School of Business, NYU, 1997.
[5] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", 1st Edition, Harcourt India Private Limited. 2001.
[6] M. H. Dunham, " Data Mining: Introductory and Advanced Topics ",1st Edition, Pearson Education (Singaphore) Pte. Ltd., 2003.
[7] G. Piateskey-Shapiro, and C. J. Matheus, "The Interestingness of Deviations", In Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 1994.
[8] S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh, "Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text ", In Proceedings of the NAACL workshop and other Lexical Resources: Applications, Extensions and Customizations, 2001.
[9] A. Silberschatz and A.Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery", In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining. 1995.
[10] B. Liu, W. Hsu, and S. Chen, " Using General Impressions to Analyse Discovered Classification Rules ", In Proceedings of the 3rd International Conférence on Knowledge Discovery and Data Mining (KDD 97), 1997.
[11] T. Kohonen, " Self-Organization and Associative Memory ", 3rd Edition, Springer, Berlin. 1993.
[12] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems ", IEEE Transactions on Knowledge and Data Engineering. V.5, No.6. 1996.
[13] B. Liu and W. Hsu, " Post Analysis of Learned Rules ", In Proceedings of the 13th National Conférence on AI(AAAI'96), 1996.
[14] S. Marsland, " On-Line Novelty Detection Through Self-Organization, with Application to Robotics ", Ph.D. Thesis, Department of Computer Science, University of Manchester, 2001.
[15] N. Japkowicz , C. Myers, and M. Gluck, " A Novelty Detection Approach to Classification", In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995.
[16] S. Roberts, and L. Tarassenko, "A Probabilistic Resource Allocation Network for Novelty Detection", In Neural Computation, 6(2), 1994
[17] [17] A. Ypma, and R. Duin, "Novelty Detection Using Self-Organizing Maps", In Progress in Connectionist-Based Information Systems. Volume 2, 1997.
[18] http://kdd.ics.uci.edu/
[19] http://www.comp.nus.edu.sg/~dm2/p_download.html