A Unified Approach for Discovery of Interesting Association Rules in Medical Databases

Harleen Kaur[1], Siri Krishan Wasan[1], Ahmed Sultan Al-Hegami[2], and Vasudha Bhatnagar[3]

[1] Department of Mathematics, Jamia Millia Islamia, New Delhi-110 025, India
harleen_k1@rediffmail.com, skwasan@yahoo.com
[2] Department of Computer Science, Sana'a University, Sana'a, Yemen
ahmed_s_gamil@yahoo.com
[3] Department of Computer Science, University of Delhi, New Delhi-110 007, India
vbhatnagar@cs.du.ac.in

Abstract. Association rule discovery is an important technique for mining knowledge from large databases. Data mining researchers have studied subjective measures of interestingness to reduce the volume of discovered rules and to improve the overall efficiency of the knowledge discovery in databases process (KDD). The objective of this paper is to provide a framework that uses subjective measures of interestingness to discover interesting patterns from association rules algorithms. The framework works in an environment where the medical databases are evolving with time. In this paper we consider a unified approach to quantify interestingness of association rules. We believe that the expert mining can provide a basis for determining user threshold which will ultimately help us in finding interesting rules. The framework is tested on public datasets in medical domain and results are promising.

Keywords: Knowledge discovery in databases (KDD), data mining, association rule, domain knowledge, interestingness, medical databases.

1 Introduction

The vast search space of hidden patterns in the massive databases is a challenge for the KDD community [19]. However, a vast majority of these patterns are pruned by the objective measures such as 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 [3].

Despite massive reduction of search space by employing suitable score functions 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 [1,2,4,5,6,7] and providing good measures of interestingness [8-18].,

Commonly used techniques to discover interesting patterns in most KDD endeavors are partially effective unless combined with subjective measures of interestingness

P. Perner (Ed.): ICDM 2006, LNAI 4065, pp. 53 – 63, 2006.

[22,24,25,26]. Subjective measures quantify interestingness based on the user understandability of the domain. Capturing the user subjectivity in dynamic environment requires a great deal of knowledge about databases, the application domain and the user's interests at a particular time [21,22,23]. Therefore, it is difficult for the user to analyze the discovered patterns and to identify those patterns that are interesting from his/her point of view.

In this paper we introduce a unified approach to quantify interestingness of association rules. The user domain knowledge is provided in terms of expert mining rules. Such expert rules are needed in order to capture the subjectivity of medical experts. The paper introduces a technique that efficiently mines the expert knowledge to form a constraint to the proposed approach. We believe expert mining can provide a basis for determining user threshold which will ultimately help as in finding interesting rules.

2 Related Works

Most existing approaches of finding subjectively interesting association rules ask the user to explicitly specify what types of rules are interesting and uninteresting. In template-based approach, the user specifies interesting and uninteresting association rules using templates [14,15,16]. A template describes a set of rules in terms of items occurring in the conditional and the consequent parts. The system then retrieves the matching rules from the set of discovered rules.

There are various techniques for analyzing the subjective interestingness of classification rules [10,11,13,14]. However, those techniques cannot work for analyzing association rules. Association rules require a different specification language and different ways of analyzing and ranking the rules. Padmanabhan and Tuzhilin have proposed a method of discovering unexpected patterns that considers a set of expectations or beliefs about the problem domain [14,15,16]. The method discovers unexpected patterns using these expectations to seed the search for patterns in data that contradict the beliefs. However, this method is generally not as efficient and flexible as our post-analysis method unless the user can specify his or her beliefs or expectations about the domain completely beforehand, which is very difficult, if not impossible [9]. Typically, the user must interact with the system to provide a more complete set of expectations and find more interesting rules. The proposed post-anlaysis method facilitates user interaction because of its efficiency. Padmanabhan and Tuzhilin's approach also does not handle user's rough or vague feelings, but only precise knowledge. User's vague feelings are important for identifying interesting rules because such forms of knowledge are almost as important as precise knowledge.

However, all works stated in the literature are generally not flexible to handle the evolving nature of data as the post-analysis method, unless the user can freely specify his or her beliefs or his/her background knowledge about the domain, which is very difficult. Liu et al. [9,10,11] proposed a post analysis method that considers vague feelings for identifying interesting rules. However, the work does not consider the degree of interestingness and the fact that the user background knowledge changes with the time.

3 The Unified Approach to Quantify Interestingness of Association Rules

An association rule is of the form: Å → C where Å denotes an antecedent and C denotes a consequent. Both Å and C are considered as a set of conjuncts of the form $c_1,c_2,\ldots,c_k$. The conjunct c_j is of the form < A = I>, where A is an item name (attribute), Dom (A) is the domain of A, and I (value) ∈ Dom (A).

Given a dataset D collected over the time $[t_0,t_1,t_2,\ldots t_n]$. At each time instance t_j, an incremental dataset D_j , j ϵ {j,…,n}, is collected and stored in D. The incremental D_i is subjected to the mining algorithm resulting in the discovery of set of rules (model) $\{R_i\}$. The proposed framework process interesting rules from the discovered rules.

Data-mining research has shown that we can measure a rule's interestingness using both objective and subjective measures [7-18]. To the end user, rules are interesting if:

(i) The rules contradict the user's existing knowledge or expectations (Unexpected).
(ii) Users can do something with them and benefit (Actionable).
(iii) They add knowledge to the user prior knowledge (Novel).

Although novelty, actionability and unexpectedness of the discovered knowledge are the basis of the subjective measures, their theoretical treatment still remains a challenging task [13,20,25]. Actionability is the key concept in most applications. Actionable rules let users do their jobs better by taking some specific actions in response to the discovered knowledge. Actionability, however, is an elusive concept because it is not feasible to know the space of all rules and the actions to be attached to them. Actionability is therefore is implicitly captured by novelty and unexpectedness [25].

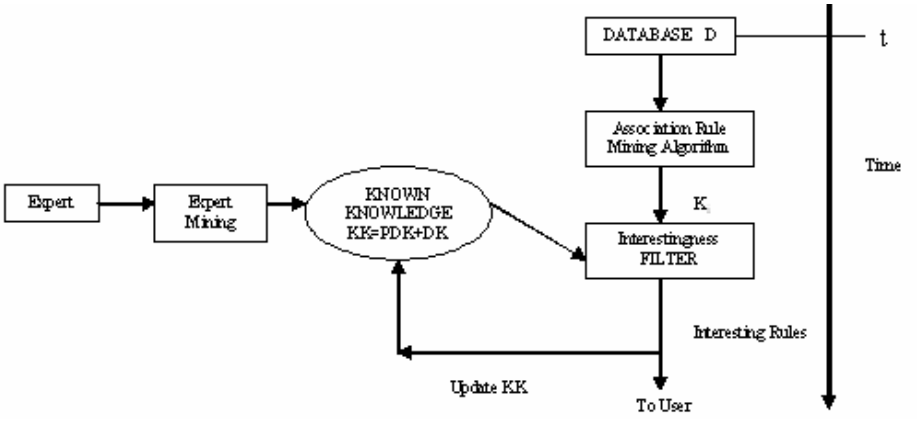

Fig. 1. Interestingness as post analysis filter for KDD process

In this work we introduce a comprehensive interestingness measure that quantifies the unexpectedness and novelty by involving the user background knowledge and the previously discovered knowledge. The framework computes the deviation of discovered rules with respect to the domain knowledge and previously discovered rules. Subsequently the user determines a certain threshold value to report interesting rules. The general architecture of the proposed framework is shown in Fig. 1.

At time t_i, database D_i is subjected to the association rule mining algorithm, resulting into discovery of knowledge K_i. The proposed interestingness filter processes K_i, in the light of knowledge extracted from expert and the previously discovered knowledge *(known knowledge)* to deliver rules that are of real interest to the user.

3.1 Deviation at Lowest Level

Degree of deviation at the lowest level represents the deviation between conjuncts. The deviation between a conjunct c_i and conjuncts c_j is computed on the basis of the result of comparison between the items of the two conjuncts.

Definition 1
Let c_1 and c_2 be two conjuncts $(A_1 = I_1)$ and $(A_2 = I_2)$ respectively. The deviation of c_1 with respect to c_2 is defined as a Boolean function as follows:

$$\Delta(c_1,c_2) = \begin{cases} 0, & \text{if } A_1 = A_2, \text{ and } I_1 = I_2 \quad \text{(Identical items).} \\ 1, & \text{if } A_1 = A_2, \text{ and } I_1 \neq I_2 \quad \text{(Different items).} \end{cases}$$

The possibilities of deviation at the lowest level as defined in Definition 1 has deviation degree 0 which indicates no deviation exists between the two conjuncts and deviation degree 1 which indicates different conjuncts.

3.2 Deviation at Intermediate Level

This type of deviation represents the deviation between the set of conjuncts. Such deviation denoted by $\Psi(S_1,S_2)$ is obtained by computing the deviation at the lowest level and subsequently combining it to compute the deviation at intermediate level. The following definition is the basis of computation of deviation at intermediate level.

Definition 2
Let S_1 and S_2 be two sets of conjuncts, we compute the deviation at intermediate level denoted by $\Psi(S_1, S_2)$ as follows:

$$\Psi(S_1,S_2) = \begin{cases} 0, & \text{iff } |S_1| = |S_2|, \forall\, c_i \in S_1, \exists\, c_j \in S_2 \\ & \quad \text{such that } \Delta(c_i,c_j) = 0 \quad \text{(Identical sets).} \\ 1, & \forall c_i \in S_1, \neg\exists\, c_j \in S_2 \quad \text{such that } \Delta(c_i,c_j) = 1 \quad \text{(Totally different).} \\ \beta, & \text{otherwise} \quad \text{(Intermediate).} \end{cases}$$

where $\beta = \dfrac{1}{|S_1|} \sum_{c_i \in S_1,\, c_j \in S_2} \min\ \Delta(c_i, c_j)$

As per Definition 2, $\Psi(S_1, S_2) = 0$ indicates that S_1 and S_2 are identical, $\Psi(S_1,S_2) = 1$ indicates the extreme deviation and the computed value of β, quantifies an intermediate degree of deviation. The value of β is computed as a linear combination of the minimum deviation at the lowest level that represents each conjunct of the S_1 with respect to S_2 divided by the number of conjuncts of S_1.

4 Interestingness of Discovered Knowledge

Having obtained the deviation at lowest and the intermediate level, the deviation at rule level (high level) is to be evaluated as both antecedents and consequents of rules are considered to be sets of conjuncts. The computation of deviation at high level is performed against the rules extracted from experts as well as the rules discovered earlier. The interestingness of a rule is therefore, obtained by comparing the deviation at the highest level (rule level) with respect the user given threshold value. A rule is considered to be interesting if its deviation at the high level exceeds a user threshold value.

Interestingness of a rule R_1 with respect to another rule R_2 is calculated as follows:

Definition 3
Let r: $\mathring{A}_r \rightarrow C_r$ be a rule whose interestingness is to be computed with respect to the rule set R. Then

$$I_r^R = \begin{cases} 0 & \text{if } \Psi(A_r, A_s) = 0 \ \& \ \Psi(C_r, C_s) = 0 \\ (\min_{S \in R} \Psi(A_r, A_s) + \Psi(C_r, C_s))/2 & \text{if } \Psi(A_r, A_s)) \geq \Psi(C_r, C_s) \\ (\Psi(A_r, A_s) + \min_{S \in R} (\Psi(C_r, C_s))/2 & \text{if } \Psi(A_r, A_s)) < \Psi(C_r, C_s) \\ 1 & \text{if } \Psi(A_r, A_s) = 1 \ \& \ \Psi(C_r, C_s) = 1 \end{cases}$$

As per Definition 3, $I_r^R = 0$ indicates that R_1 and R_2 are identical, $I_r^R = 1$ indicates the extreme deviation between R_1 and R_2. $(\min_{s \in R} \Psi(A_r,A_s) + \Psi(C_r,C_s))/2$ and $(\Psi(A_r,A_s) + \min_{s \in R} \Psi(C_r,C_s))/2$ indicates the intermediate degree of deviation of R_1 with respect to R_2. The user specifies the threshold to select interesting rules based on the computation of I_r^R.

After rule interestingness is computed, we have to decide either the rule is interesting or simply a deviation of an existing rule. Whether a rule is interesting or not depends on the user feeling about the domain, which is determined by a certain threshold value. The following definition is the basis of determining interesting rules.

Definition 4
Let R_1: $\mathring{A}_1 \rightarrow C_1$ and R_2: $\mathring{A}_2 \rightarrow C_2$ be two association rules. R_1 is considered interesting with respect to R_2, if $I_{R1}^{R2} > \Phi$, where Φ is a user threshold value, otherwise it is considered conforming rule.

As per Definition 4, the computed value I_{R1}^{R2} which indicates the interestingness of R_1 with respect to R_2 is compared against the user threshold value Φ to determine either R_1 is interesting with respect to R_2 or otherwise. The R_1 is interesting if its deviation with respect to R_2 exceeds Φ.

5 Expert Mining Using Mathematical Techniques

Most Association rule algorithms employ support-confidence threshold to exclude uninteresting rules but in medical data mining, many rules satisfying minimum confidence and minimum support may not be interesting in view of expert's experience of critical cases. It is only the user (medical expert) who can judge if the rule is interesting or not. The judgment being subjective, will vary from expert to expert.

Traditionally, medical expert system extract knowledge using IF-THEN diagnostic rules, where as data mining algorithms use large databases to discover a set of rules. Machine learning techniques too rely on available databases. In case of medical databases, it is possible that there are many missing or incomplete records. On the other hand a medical expert because of his limited experience may arrive at incorrect rule. Therefore, it is desirable to compare rules generated by data mining algorithms with rules generated by experts. Subsequently, contradictions can be identified and eliminated to discover interesting rules.

We may extract rules from medical experts using mathematical techniques. Kovalerschuk et al. have applied monotonicity of Boolean functions in the breast cancer problem by evaluating calcifications in a mammogram [27]. Suppose we identify n attributes say $x_1, x_2, x_3.....x_n$ to diagnose a particular disease D. Without loss of generality, we assume these attributes take binary values yes or no i.e. 1 or 0 then there are 2^n combinations of these attributes. We can extract rules by interviewing medical experts on these 2^n combinations of the values of the attributes. By using monotonicity in some form on these 2^n vectors, we may minimize the number of questions. One simple way of defining monotonicity is as follows:

$$(x_1, x_2, x_3.....x_n) \leq (y_1, y_2, y_3.....y_n) \quad \text{iff } x_i \leq y_i$$

Now questions to expert will depend on answer to the previous question. Chain of monotonic values of $(x_1, x_2, x_3.....x_n)$ represents a case using Hansel chain [28].

6 Implementation and Experimentation

The proposed approach is implemented and tested on several public medical datasets available at http://kdd.ics.uci.edu using C programming language. The datasets are partitioned into three groups representing instances arrived at time T_1, T_2 and T_3 respectively. The rules are generated using WEKA-associate [29] for each partition of the datasets, with 0.1% and 1% to indicate minimum confidence and minimum support respectively. Subsequently, their interestingness is quantified using the proposed framework. Based on the specified threshold the rules are categorized either as interesting or conforming (Definition 4).

6.1 Experiment I

The objective of the first experiment is to show the effectiveness of the approach in reducing the number of discovered rules. It is expected that the number of discovered rules that are interesting keeps on decreasing over the time. We work with five datasets and assume that the interestingness threshold value (Φ) = 0.6. The values in the third column of Table 1 represent the number of rules discovered, using WEKA, at a given partition and the values in the fourth column represent the interesting rules discovered by our approach. It is observed that the number of interesting rules decreases in contrast to the number of conforming rules which increases as expected. Intuitively, the

Table 1. The discovered medical rules at time T_1, T_2, and T_3

Dataset	Time	Discovered AR's	Interesting rules	Conforming rules
Lymph	T_1	32000	18230	13770
	T_2	28562	12003	16559
	T_3	26781	2010	24771
Breast	T_1	802	320	482
	T_2	725	180	545
	T_3	540	73	467
Hepatitis	T_1	1207	800	407
	T_2	980	430	550
	T_3	626	228	398
Heart	T_1	987	564	423
	T_2	566	320	246
	T_3	207	118	89
Sick	T_1	4502	2876	1635
	T_2	2709	1078	1631
	T_3	986	401	585

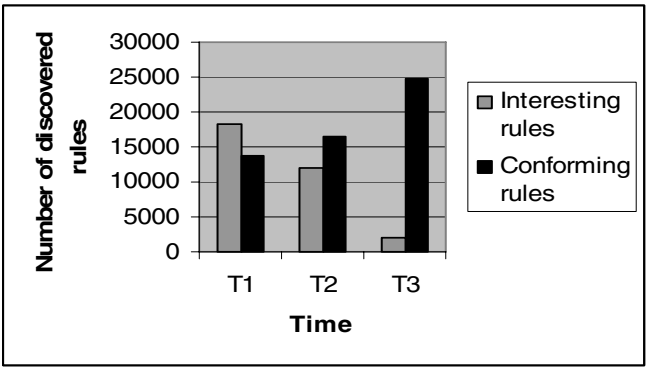

Fig. 2. Graphical representation of discovered rules of Lymph dataset

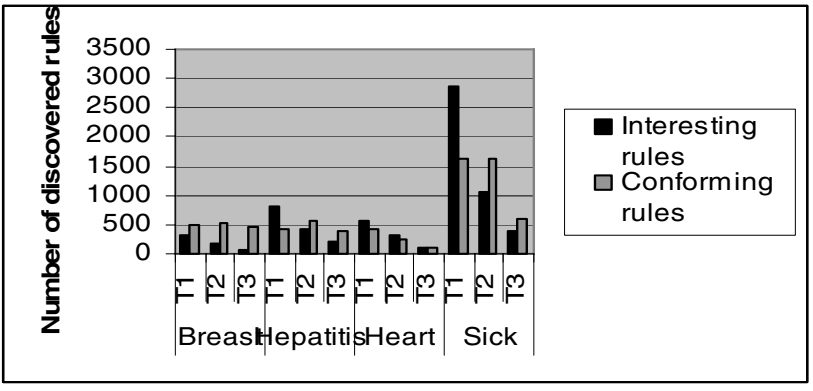

Fig. 3. Graphical representation of discovered rules of different datasets

Table 2. Discovered rules at time T_1, T_2 and T_3 for different (Φ)

Interesting Degree (Φ)	Time	Discovered Rules	Interesting	Conforming
Φ=0.9	T_1	1207	291	913
	T_2	980	160	820
	T_3	626	119	507
Φ=0.8	T_1	1207	311	896
	T_2	980	259	721
	T_3	626	156	470
Φ=0.7	T_1	1207	417	790
	T_2	980	388	592
	T_3	626	214	412
Φ=0.6	T_1	1207	800	407
	T_2	980	430	550
	T_3	626	228	398
Φ=0.5	T_1	1207	976	231
	T_2	980	530	450
	T_3	626	324	302
Φ=0.4	T_1	1207	1016	191
	T_2	980	860	120
	T_3	626	520	106
Φ=0.3	T_1	1207	1103	104
	T_2	980	923	57
	T_3	626	602	24

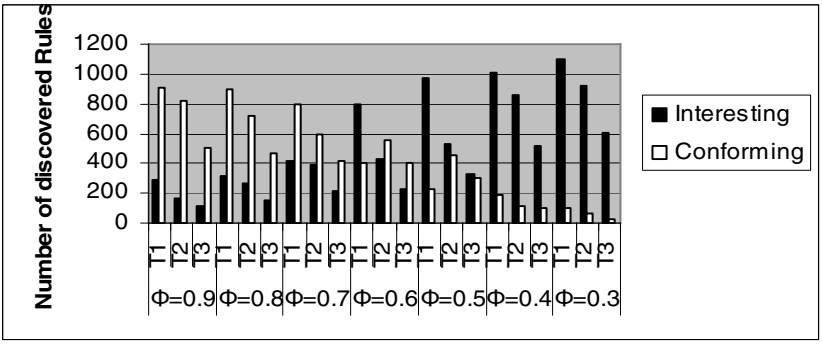

Fig. 4. Graphical representation of discovered rules

interesting rules discovered at time T_1 become known knowledge at time T_2 and hence no more interesting. The conforming rules are shown in the rightmost column of Table 1. Figures 2 and 3 shows the graphical representation of Table 1.

6.2 Experiment II

The second experiment was performed using 'Hepatitis' dataset to study the effectiveness of interestingness threshold (Φ) on the number of discovered rules. It is expected that as the interestingness threshold value (Φ) decreases, the number of rules increases. Intuitively, a higher value of Φ indicates that the user background knowledge about the domain is high and therefore number of interesting rules is reduced. In contrast, a lower value of Φ indicates that the user background knowledge about the domain is low and therefore number of interesting rules is increased. Table 2 shows the result of this experiment. Fig. 4 shows the graphical representation of the results.

7 Conclusions

In this paper, we proposed framework to quantify the interestingness of association rules in evolving medical databases. The approach is post-analysis filter that is used in analysis stage of KDD process. It is based on computation of the deviation of the currently discovered association rules with respect to expert rules and previously discovered knowledge. The user subjectivity is captured the by constructing the expert rules. The framework is implemented and evaluated using five medical datasets and has shown encouraging results.

Currently we are trying to integrate the framework into the Apriori algorithm (mining algorithm), thus using it in the mining stage of the KDD process.

References

1. Han, J. and Kamber, M.: Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers, (2001)
2. Dunham M. H.: Data Mining: Introductory and Advanced Topics. 1st Edition Pearson ygEducation (Singapore) Pte. Ltd. (2003)

3. Hand, D., Mannila, H. and Smyth, P.: Principles of Data Mining, Prentice-Hall of India Private Limited, India, (2001)
4. Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Adaptive Constraint Pushing in Frequent Pattern Mining. In Proceedings of the 17th European Conference on PAKDD03 (2003)
5. Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: ExAMiner: Optimized Level-wise Frequent pattern Mining with Monotone Constraints. In Proceedings of the 3rd International Conference on Data Mining (ICDM03) (2003)
6. Bronchi, F., Giannotti, F., Mazzanti, A., Pedreschi, D.: Exante: Anticipated Data Reduction in Constrained Pattern Mining. In Proceedings of the 7th PAKDD03 (2003)
7. Freitas, A. A.: On Rule Interestingness Measures. Knowledge-Based Systems. 12:309-315 (1999)
8. Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A. I.: Finding Interesting Rules from Large Sets of Discovered Association Rules. In Proceedings of the 3rd International Conference on Information and Knowledge Management. Gaithersburg, Maryland (1994)
9. Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems (2000)
10. Liu, B., Hsu, W.: Post Analysis of Learned Rules. In Proceedings of the 13th National Conference on AI (AAAI'96) (1996)
11. Liu, B., Hsu, W., Lee, H-Y., Mum, L-F.: Tuple-Level Analysis for Identification of Interesting Rules. In Technical Report TRA5/95, SoC. National University of Singapore, Singapore (1996)
12. Liu, B., Hsu, W.: Finding Interesting Patterns Using User Expectations. DISCS Technical Report (1995)
13. Liu, B., Hsu, W., Chen, S.: Using General Impressions to Analyze Discovered Classification Rules. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data mining (KDD 97) (1997)
14. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a Measure of Interestingness in Knowledge Discovery. Working paper # IS-97-. Dept. of Information Systems, Stern School of Business, NYU (1997)
15. Padmanabhan, B., Tuzhilin, A.: A Belief-Driven Method for Discovering Unexpected Patterns. KDD-98 (1998)
16. Padmanabhan, B., Tuzhilin, A.: Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns. KDD-2000 (2000)
17. Piatetsky-Shapiro, G., Matheus, C. J.: The Interestingness of Deviations. In Proceedings of AAAI Workshop on Knowledge Discovery in Databases (1994)
18. Piatetsky-Shapiro, G.: Discovery, Analysis, and Presentation of Strong Rules. In Knowledge Discovery in Databases. The AAAI Press (1991)
19. Psaila, G.: Discovery of Association Rules Meta-Patterns. In Proceedings of 2nd International Conference on Data Warehousing and Knowledge Discovery (DAWAK99) (1999)
20. Agrawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases, In ACM SIGMOD Conference of Management of Data. Washington D.C., (1993)
21. Silberschatz, A., Tuzhilin, A.: On Subjective Measures of Interestingness in Knowledge Discovery. In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (1995)
22. Silberschatz, A., Tuzhilin, A.: What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Trans. and Data Engineering. V.5, no.6 (1996)

23. Suzuki, E., Kodratoff, Y.: Discovery of Surprising Exception Rules Based on Intensity of Implication. In Proceedings of the 2nd European Symposium, PKDD98, Lecture Notes in Artificial Intelligence (1998)
24. Liu, B., Hsu, W., Chen, S., and Ma Y.: Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems (2000)
25. Al-Hegami, A. S., Bhatnagar, V. and Kumar, N.: Novelty Framework for Knowledge Discovery in Databases. In Proceedings of the 6th International Conference on Data warehousing and Knowledge Discovery (DaWak 2004). Zaragoza, Spain, pp 48-55 (2004)
26. Bhatnagar, V., Al-Hegami, A. S. and Kumar, N.: Novelty as a Measure of Interestingness in Knowledge Discovery. In International Journal of Information Technology, Volume 2, Number 1 (2005)
27. Kovalerchuk, B., Triantaphyllou, E., Despande, A. and Vtyaev, E.: Interactive Learning of Monotone Boolean Function. Information Sciences, 94 (1-4):87-118 (1996)
28. Hansel, G.: Sur le nombre des functions Boolenes Monotones den variables. C.R. Acad. Sci. Paris, 262(20):1088-1090 (in French) (1966)
29. Witten, I.H. and Frank, E.: Data Mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco (2000)