Mining Interesting Rules in Bank Loans Data

Nazlı İkizler and H.Altay Güvenir

Bilkent University,
Department of Computer Engineering,
Bilkent, Ankara 06533, Turkey
{inazli,guvenir}@cs.bilkent.edu.tr

ABSTRACT

Most of the data mining algorithms produce a long list of rules in which it is up to user to find the ones that are really important and profitable. An automated process is needed in order to eliminate this undesired situation. What we may call a second level of data mining is this process of finding interesting rules among already generated ones. In this paper, rule interestingness measures are discussed and a new rule selection mechanism is introduced. This new method has been applied to learning interesting rules for the evaluation of bank loan applications. C4.5, a decision tree classifier, is used in generating the rules of the domain. Experimental results show that our method is effective in finding interesting rules.

I. INTRODUCTION

A data mining technique usually generates a large amount of patterns and rules. However, most of these patterns are not interesting from a user's point of view. Beneficial and interesting rules should be selected among those generated rules. This selection process is what we may call a second level of data mining; *mining among rules*. Although there are several studies on this subject, there is not a standard or universally best measure to decide which rule is interesting or which one is not, yet.

A pattern is *interesting* if it is easily understood, unexpected, potentially useful and actionable, novel, or it validates some hypothesis that a user seeks to confirm [1]. There are two aspects of rule interestingness, objective and subjective aspects and this fact is what makes finding interesting rules a challenging issue. Objective, data-driven, aspect is based on statistics and structures of patterns, e.g., support, confidence, etc. On the other hand, subjective side is user-driven, it is based on user's belief in data, e.g., unexpectedness, novelty, actionability, etc.

In this study, we will be concentrating on the objective measures of interestingness. However, one should note that both aspects should be used to select interesting rules, since a data mining process mostly needs subjective evaluation of the domain.

In this paper, the generic problem of finding the interesting ones among generated rules is addressed and a case study has been carried out over a bank loan data. In this case study, rules are learned using the decision tree classifier C4.5. In Section 2, some of proposed measures of rule interestingness are introduced. Section 3 discusses how the proposed measures can be extended to take additional factors into account. A study and implementation of finding interesting rules for bank loan data is explained in Section 4. Section 5 gives the experimental results over our specific domain. Section 6 makes a concise conclusion of the study and points out future work in this area.

II. RULE INTERESTINGNESS BASICS

A classification rule is in the form $A \Rightarrow B$ where A is a conjunction of predicting attribute values and B is the predicted class. Quality of the rule is mainly evaluated by looking at three factors, namely coverage, completeness and the confidence of the rule. Coverage, denoted by $|A|$, is number of tuples satisfied by rule antecedent. Completeness of the rule is $|A\&B|/|B|$, the proportion of number of tuples of the target class covered by the rule. The rule's confidence factor, in other words predictive accuracy of the rule, is given by $|A\&B|/|A|$.

Piatetsky-Shapiro proposed three principles for rule interestingness (RI), which can be stated as follows [2]:

1. RI = 0, if $|A\&B| = |A||B|/N$ (Here N is the number of tuples. Rule interestingness is 0 if rule antecedent and rule consequent are statistically independent.)
2. RI monotonically increases with $|A\&B|$ when other parameters are fixed.
3. RI monotonically decreases with $|A|$ or $|B|$ when other parameters are fixed.

Freitas has pointed out five additional factors related to the interestingness of a rule [3]. These are disjunct size, imbalance of the class distribution, attribute costs, misclassification costs and asymmetry in classification rules. These factors are claimed to be important when designing a domain-dependent rule interestingness measure.

Small disjunct problem, as it is addressed in [3], is related to the tendency of most data mining algorithms to discover large disjuncts. This bias towards large disjuncts can eliminate very important and interesting knowledge if not treated properly. Especially, in domains where small disjuncts collectively match a large percentage of the tuples, this is an undesired situation. In order to overcome this problem, small and large disjuncts should be evaluated in different ways.

In some of the domains, the tuples belonging to one class are much more frequent than the tuples belonging to other classes. This is called imbalance of the class distribution. In such a case, the smaller the relative frequency of a minority class, the more difficult to discover rules predicting it. So, from the interestingness point of view, the rules predicting the minority class are more interesting.

Freitas has also mentioned the fact that most rule interestingness measures consider the rule antecedent as a whole. However, as he states, the interestingness of two rules having the same coarse-grained value may be different, depending on the attributes occurring in the rules antecedent. So, a good interestingness measure should consider attributes individually, according to their specific properties that may be domain dependent. Attribute costs are one of these properties. In some application domains such as medical diagnosis, different attributes might have very different costs. In such a case, a rule whose antecedent consists of less costly attributes are more interesting.

Misclassification cost is another issue, which should not be ignored when designing a good interestingness measure. Especially in domains where misclassifying a case is highly crucial, users of the domain are in need of a less risky classification system. In order to achieve such a system, misclassification costs of the produced rules should be reasonable. In other words, the smaller the misclassification cost of a rule, the more interesting it is.

The last factor that has been stated by Freitas is the asymmetry in classification rules. A rule interestingness measure is said to be symmetric with respect to the rule antecedent and the rule consequent. The reason for this is that we want to discover rules where the value of predicting attributes determine the value of the goal attribute.

Besides all these, there are several other rule interestingness measures in the literature. Some of these measures are discussed in [4]. Most of these measures depend on statistical factors such as correlation. Lately, Tan and Kumar proposed another measure, called IS, which is again derivable from statistical correlation [5]. We will revisit this measure in Section 3.

Another issue regarding the interestingness of the rules is the interaction between the rules. Currently used techniques in data mining can generate redundant patterns along with the desired ones. The generated rules can have the same semantic information, which we may call *overlapping* of functions. In order to generate really useful and interesting rule patterns, these rule interaction effects should also be taken into account by the rule selection mechanism.

III. INTERESTING RULE SELECTION

As explained in Section 2, there are several factors that should be considered while determining the interestingness of the generated rules. In this section, we discuss how these factors should be evaluated and integrated into the rule selection mechanism. For this reason, we present a step-by-step schema to produce really interesting rules.

Pruning Redundant Rules

If the data mining technique used for extracting rules produces redundant rules, this redundancy should be eliminated before calculating interestingness of rules. Here we may basically say a rule is redundant if it satisfies one of the following two conditions:

1. If there are two implications of the form $A \rightarrow C$ and A & $B \rightarrow C$, and both rules have similar confidence values, then the rule A & $B \rightarrow C$ is redundant.
2. If there are two implications $A \rightarrow C$ and $B \rightarrow C$, both have similar confidence values, then $B \rightarrow C$ is redundant if B is a subset of the conditions of A.

There may be other conditions besides the ones given above, but it is shown that most of the redundant rules are detected by these two conditions [6]. Similar confidence values mentioned in both situations are confidence values differing within a small range, e.g. 0.05. The first principle says that if the addition of one condition to the rule antecedent does not affect the confidence of a rule, then, addition of that condition is unnecessary. The second principle says that the subsets of a generated rule are redundant if they are of the similar confidence strength.

If the two conditions given above are carefully observed, it is clear that the redundant rules are not interesting from the user’s point of view, because they are mostly generated due to their relation with the more general rules. Hence, in the interesting rule selection mechanism, the first step should be pruning of these uninteresting rules.

Grouping by Coverage Values

The second step of determining the interesting rules should be grouping the pruned rules into subgroups according to their coverage values. As it is noted in Section 2, small and large disjuncts; i.e., the rules having low coverage (support) and rules having large coverage should be evaluated in different ways. For this evaluation, Holte et al. suggested that small disjuncts should be evaluated with a maximum-specificity bias, in contrast with the maximum-generality bias favouring large disjuncts [7].

Tan and Kumar also argued that a good interestingness measure should take into account the support of the pattern [5]. They have showed that their proposed IS measure can be used in the region of low support, i.e., support of 0.3, whereas using RI measure in the region of high support is preferred. Hence, in order to make small disjuncts as interesting as large disjuncts, we may take IS measure as the basic measure for rules having coverage values in the range of [0,0.3] and RI measure for rules with coverage values [0.3,1]. Here are the formulations for two measures respectively:

$$IS = \sqrt{\frac{P(A,B)P(A,B)}{P(A)P(B)}}$$

$$RI = P(A,B) - P(A)P(B)$$

In this paper, these two measures are experimented over bank-customers' loans data individually and results are presented in Section 5.

Misclassification Costs

So far, we have removed redundant rules and decided on the basic rule interestingness measure based on the coverage of the rule. Now, we can consider other factors and build our interestingness measure properly.

We have seen that misclassification costs are important for rule interestingness. A rule predicting a patient does not have a particular disease while he indeed does is very risky and misclassification cost of such a rule is very high. In domains where we cannot tolerate erroneous classifications, a rule which has a low error rate and low misclassification cost is more desirable.

In order to integrate this inverse proportion to the rule interestingness calculations, we should divide the basic rule interestingness measure by the misclassification cost of the rule, which is defined as follows:

$$\text{MisclasCost} = \sum_{j=1}^{k} \text{Prob}(j)\text{Cost}(i,j)$$

Here Prob(j) is the probability that a tuple classified by the rule has true class j, class i is the class predicted by the rule and Cost(i,j) is the cost of misclassifying a tuple with true class j as class i and k is the number of classes. Prob(j) can be calculated as follows, taking into account the effect of disjunct size:

$$\text{Prob}(j) = (1 + |\text{A}\&\sim\text{B}|)/(k + |\text{A}|)$$

Attribute Factors

In most of the interestingness measures, the rule antecedent is taken as a whole and interest in the individual attributes is ignored. To overcome this situation, we should modify our interestingness measure to take important attribute factors into account. For example, in some domains attributes may have acquirement costs. In a medical diagnosis phase, the cost of getting a computer tomography is much higher than determining the blood pressure. In such domains if the prediction accuracy of two rules are similar, the rule with low attribute costs should be preferred.

Freitas has proposed a term called *attribute usefulness* defined as the inverse of the sum of the costs of all the attributes occurring in the rule antecedent, that is:

$$\text{AttUsef} = 1 / \sum_{i=1}^{k} \text{Cost}(A_i)$$

where $\text{Cost}(A_i)$ is the cost of the *i*th attribute occurring in the rule antecedent and *k* is the number of attributes in the rule antecedent.

Besides attribute costs, there may be other factors related to attributes individually. For instance, a user may be interested in specific attributes and rules containing those attributes. In that case attributes may have different *interest weights* and the more the interest weights the more interesting the rule is. In some domains, for example, one user may be interested in finding the geographical causes of some disease, while another user is interested in finding effects of demographical features.

Since the attribute factors are mostly domain specific, data mining experts should know about these factors and should decide on the ones to include in the interestingness measure. If attributes have different costs, the basic rule interestingness measure should be multiplied with Attribute Usefulness measure. Other factors can be considered in a similar way.

Class Weights

Discovery of a particular class may be more interesting and useful for a data mining system in a specific domain. For the rule selection mechanism, the classes can be given interest weights beforehand and by this way, rules predicting classes with high interest rates can have larger interestingness values. In order to accomplish this goal, rule interestingness measure can be multiplied by the weight of the predicted class of the rule. Note that, as we mentioned above, if no weights are specified the rules discovering the minority class are of higher interest than the rules predicting the majority class.

IV. FINDING INTERESTING RULES IN BANK LOANS DATA

In this study, the rule interestingness criteria mentioned above are applied to a real world situation, where the data to be mined consists of bank customers who have borrowed some amount loan from a particular bank.

Domain Description

In our experimentation domain, we are trying to predict whether a customer will pay the bank loan back or not by looking at some information about the customer. The cases where the applicants have paid their loan back are represented by class 0, while the others are class 1.

There are 13 attributes in our domain; 7 of these attributes are metric and 6 of them are non-metric (categorical). The list of attributes and their types is given in Table 1.

Table 1. Attributes of the domain.

Type of Attributes	Name of Attributes
Metric	Age, Years in the same address, Years in the same job, Total income, Credit amount, Credit/(Term*income), Fixed Term

Categorical	Married, Have children, Have car, Have house, Occupation, Home Place

Training data of our data set consists of 1300 cases and 391 of these cases are of class 1 and remaining 909 cases are of class 0. Test data consisted of 143 cases. The decision tree induction algorithm C4.5 has been used in order to generate rules of the experimentation domain. As a result, a total of 184 rules are generated by the standard. Then these rules are evaluated with the rule interestingness criteria.

An example rule generated by the C4.5 algorithm is given below:

Rule 1: occupation = other
same_address <= 24
term <= 5
-> class 0 [96.8%]

where 1 represents the number of the rule, each line represents the conditions of the rule and the last line shows the predicted class along with the confidence of the rule in the whole data set (both in test and training data).

Experimental Details

In our domain, the attributes have no acquirement costs since they are mostly demographic information. Also, the attributes are not given weights, since all attributes are of the same interest. Although we are interested in predicting customers who are not likely to pay the money back (class 1 instances), we do not assign any weights to classes either. Since class 1 is the minority class in the data set, and rules predicting the minority class are already counted as more interesting, assignment of weights to classes is not essential in our example domain.

It is obvious that classifying a customer as he will pay the money back, while indeed he will not is much more costly than classifying a customer as he will not pay the money back, while he will. In the former case, the bank will lose money whereas in the latter case it will lose profit. So the assigned misclassification costs should be different in different cases and the costs assigned in our experimentation is listed below in Table 2:

Table 2. Misclassification costs of the domain

	Predicted	
Actual	Class 0	Class 1
Class 0	0	2
Class 1	20	0

V. EXPERIMENTAL RESULTS

Interestingness measures presented in Section 3 are applied to the loan applications domain. Before we eliminate redundancy, first ten rules with their interestingness values in a sorted decreasing order are presented Table 3. Here, all the interestingness values are calculated by taking the RI measure as the basic measure.

Table 3. First ten interesting rules before pruning.

Rule Antecedent	**Class**	**Coverage**	**Confidence**	**Interestingness**
Rule 5:credit_term_income <= 0.964258, term <= 5	0	263	97.2	0.125879
Rule 108:age <= 25, have_child = 0 Have_house = 0, credit_amount > 1073	1	28	90.9	0.107183

credit_term_income <= 0.256667, term > 8				
Rule 86:credit_amount > 1073, term > 8	1	451	54.6	0.104145
Rule 2:same_address <= 24, term <= 5	0	288	96.3	0.097437
Rule 7:same_address <= 24, term > 2, term <= 5	0	288	97.7	0.097437
Rule 37:have_car = 0, credit_amount > 619, home = BOLU	1	18	92.6	0.096817
Rule 101:have_child = 1, total_income<=1604,term>8,home=BOLU	1	18	92.6	0.096817
Rule 61:married = 1, have_car = 0, Home = BOLU	1	17	92.2	0.086866
Rule 182: credit_term_income <= 0.18638 Term > 8, occupation = self_occupied	1	86	76.5	0.086866
Rule 139: total_income > 612, Credit_amount > 1073, term > 8	1	331	54.8	0.078665

As it can be seen from Table 3, there are rules that are interacting with each other and user who sees these rules will not be much surprised, because most of the rules capture the same semantic meaning more or less. For example, Rule 2's conditions are seen in Rule 7 and some of the rules in the rest of the list. In this case, redundancy may exhaust the user, and we need a way to eliminate this situation. In Table 4, the first ten interesting rules after the rule set has been pruned are shown.

Table 4. First ten interesting rules after pruning.

Rule Antecedent	**Class**	**Coverage**	**Confidence**	**Interestingness**
Rule 5:credit_term_income <= 0.964258, term <= 5	0	263	97.2	0.125879
Rule 108:age <= 25, have_child = 0 have_house = 0, credit_amount > 1073 credit_term_income <= 0.256667, term > 8	1	28	90.9	0.107183
Rule 86:credit_amount > 1073, term > 8	1	451	54.6	0.104145
Rule 37:have_car = 0, credit_amount > 619 Home = BOLU	1	18	92.6	0.096817
Rule 101:have_child = 1, total_income <= 1604, term > 8, home = BOLU	1	18	92.6	0.096817
Rule 61:married=1, have_car= 0,home = BOLU	1	17	92.2	0.086866
Rule 182:credit_term_income <= 0.186385, term > 8, occupation = self_occupied	1	86	76.5	0.081106
Rule 94:credit_amount <= 2095, term > 8, Occupation = self_occupied	1	73	78.2	0.078387
Rule 181:have_car = 0, term > 8, Occupation = self_occupied	1	55	80.9	0.074192
Rule 22:term > 5, occupation = self_occupied	1	123	69.1	0.071436

After pruning redundant rules rules, 139 rules are left, out of 184 rules. It is clear from Table 4 that, most of the redundancy is removed. Note that, the rules compared for redundancy have the similar confidence strength. Here, a difference amount of 0.05 is used, however, if needed, this pruning ratio can be made smaller or larger.

In Table 4, one important observation can be that the most interesting rule is the one belonging to class 0 and also in the rest of the list, rules belonging to this class are in higher levels. This is due to coverage grouping ratio of the interesting rule selection mechanism. In the first run, all rules were evaluated with RI measure. If we put a coverage threshold 0.3 and evaluate rules having coverage ratios more than this threshold with RI measure and rules with lower coverage ratios with IS measure, resulting first ten rules are shown in Table 5.

Table 5. Rule coverage ratio 0.3 for RI and IS measures.

Rule Antecedent	Class	Coverage	Confidence	Interestingness
Rule 37:have_car = 0, credit_amount > 619 Home = BOLU	1	18	92.6	2.145596
Rule 101:have_child = 1, total_income <= 1604, term > 8, home = BOLU	1	18	92.6	2.145596
Rule 61:married = 1, have_car = 0, home = BOLU	1	17	92.2	1.980887
Rule 108:age <= 25, have_child = 0 have_house = 0, credit_amount > 1073 credit_term_income <= 0.256667, term > 8	1	28	90.9	1.935342
Rule 112:have_child = 0, have_house = 0, Credit_amount <= 1894, term > 8, Occupation = self_occupied	1	15	91.2	1.664854
Rule 5:credit_term_income <= 0.964258, term <= 5	0	263	97.2	1.165264
Rule 38:credit_amount > 619, home = ICEL	1	26	85.6	1.110818
Rule 79:age > 26, have_child = 0, have_car = 1, have_house = 0, term > 8, home = ISTANBUL	1	11	88.2	1.090238
Rule 117:have_child = 0, have_car = 1, have_house = 0, credit_amount > 1073,term> 8, home = ISTANBUL	1	11	88.2	1.090238
Rule 181:have_car = 0, term > 8, Occupation = self_occupied	1	55	80.9	1.014917

It can be observed from Table 5 that our claim in Section 3 is true. By employing IS measure for the rules having low coverage values, and RI for rules with high coverage values, the minority class rules become more interesting (Rules predicting class 1 rise up in the interestingness list). Further experimentation can help us to adjust this coverage ratio more properly.

The algorithm runs in O(n*m) time where n stands for the number of cases in the data set and m is the number of rules in the rule set. It took 693msec for one run of our experimentation to complete in a 60 MHz, SparcSTATION 20 machine.

VI. CONCLUSIONS AND FUTURE WORK

In this study, we investigated the aspects of rule interestingness and discussed some criteria for deciding on the best rule interestingness measure in the domain. An interesting rule selection mechanism is proposed by combining those criteria in an effective manner. The initial experimentation results of the rule selection in a bank customers' domain are presented.

By the initial results, we see that pruning is an important part of the rule interestingness and overlapping rule cases should be considered properly for the sake of novelty principle of rule interestingness. Secondly, we see that, in domains where the rules predicting the minority class are more important, coverage (support) values of rules should be taken into account and basic rule interestingness measure should be decided according to those values. It is shown that using IS measure with rules having low support and using RI measure with the rules having large support brings out promising results.

For future work, different sizes of rule sets can be tested using our rule selection mechanism. As stated above, different thresholds can be experimented and best results can be obtained. Attributes and classes can be given weights and results can be compared. While determining misclassification costs, different characteristics of the data can be evaluated and integrated to misclassification costs. In any case, the most important action to take will be showing the obtained results to users of the domain and get feedback about the rules' real interestingness values.

REFERENCES

[1] J. Han and M. Kamber "*Data Mining: Concepts and Techniques*". Academic Press, pp.27, 2001.

[2] G. Piatetsky-Shapiro, "*Discovery, Analysis and Presentation of Strong Rules*". Knowledge Discovery in Databases, AAAI, 1991, pp.229.

[3] A.A. Freitas "*On rule interestingness measures*". Knowledge-Based Systems, 12(5-6):pp. 309--315, October 1999.

[4] R.J. Hilderman and H.J. Hamilton "*Knowledge discovery and interestingness measures: A survey*". Technical Report CS 99-04, Department of Computer Science, University of Regina, October 1999.

[5] P. Tan and V. Kumar, *"Interestingness Measures for Association Patterns: A Perspective"*. Technical Report # TR00-036, Department of Computer Science, University of Minnesota, 2000.

[6] D.Shah, L.V.S. Lakshmanan, K. Ramamritham and S. Sudarshan, *"Interestingness and Pruning of Mined Patterns"*. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), May, 1999.

[7] R.C. Holte, L.E. Acker, B.W. Porter *"Concept Learning and the Problem of Small Disjuncts"*, Proceedings of the International Joint Conference on AI (IJCAI-89), pp. 813-818.