Analysis and Presentation of Interesting Rules

Türker Yılmaz and H. Altay Güvenir

Bilkent University
Department of Computer Engineering
06533 Bilkent, Ankara, Turkey
Tel: +90 (312) 266 41 33, **Fax:** +90 (312) 266 41 26
e-mail: {yturker, guvenir}@cs.bilkent.edu.tr

ABSTRACT. Finding interesting rules from a large rule set is one of the most important aims of data mining research. In this paper, we present an algorithm and its application to finding and presenting interesting rules from a large set of previously discovered rules. Our domain of experimentation is the differential diagnosis of erythemato-squamous diseases. The application domain contains records of patients with known diagnosis.
Keywords: Data mining, Rule interestingness, Rule surprisingness.

I INTRODUCTION

It is a necessity for the discovered rules to be interesting since interesting rules may reveal many hidden and beneficial knowledge for a domain. These rules; both interesting and uninteresting, are generally defined in the form of "if-then" rules. The term *interestingness* is arguably related to the properties of surprisingness (unexpectedness), usefulness and novelty of the rule [1]. Generally, the assessment of the interestingness of discovered rules consists of objective and subjective part. In the particular case of surprisingness, it is argued that this property of the discovered rules is subjective [2]. Objective measures of rule surprisingness have one important advantage: objectiveness is strongly related to domain independence, while subjectiveness is strongly related to domain dependence. Therefore, objective measures of rule surprisingness are more generic than subjective measures.

In this paper we propose an algorithm to extract interesting rules from a large set of previously discovered rules, by using both objective and subjective objective measures. Although only a little portion of these measures that are used in calculations of the interesting ones are subjectively defined, we can say that this little subjectiveness makes almost whole approach subjective. This means that subjective measures are more effective for the determination of interesting rules.

The rest of this paper is organized as follows. Section II presents a review of several rule interestingness criteria. Section III describes the procedure followed and the algorithm for finding interesting rules. The test application and the domain parameters are explained in Section IV. In Section V, results of the algorithm are presented. Finally, Section VI concludes the paper.

II RELATED WORK

II.I RULE INTERESTINGNESS PRINCIPLES

We can express a classification rule in the form of A⇒B, where A is the conjunction of the predicting attribute values and B is the predicted class for that rule. It should be noted that, A and B can be a composite representation of many attributes. The quality of a rule is affected by three factors:

- Coverage: $|A|$ is the number of cases satisfied by the rule's antecedent.
- Completeness: $|A\&B| \,/\, |B|$ is the proportion of cases of the target class covered by the rule
- Confidence factor: $|A\&B| \,/\, |A|$ is the predictive accuracy.

Piatetsky and Shapiro, have proposed three principles for rule interestingness (RI) measures as the following [3]:

- A rule is *not* interesting if $|A\&B| = |A|\,|B|\,/N$ which means RI=0, if the antecedent and the consequent of the rule are independent in a statistical manner.
- Rule interestingness monotonically increases with the $|A\&B|$ when other parameters are fixed. This means that for fixed $|A|$ and for fixed $|B|$, rule interestingness monotonically increases as $|A\&B|$ increase. In terms of the previously mentioned rule quality factors, for fixed $|A|$ and fixed $|B|$, the confidence factor and the completeness of the rule monotonically increase with $|A\&B|$ and as these factors get higher the interestingness of the rule increases.
- Rule Interestingness monotonically decreases as $|A|$ or $|B|$ increases when other parameters are fixed. This means that for fixed $|A|$ and fixed $|A\&B|$, rule interestingness monotonically decreases as $|B|$ increases. For fixed $|B|$ and $|A\&B|$ (which implies a fixed rule completeness), rule interestingness monotonically decreases as $|A|$ increases. As the coverage increase, the confidence factor gets smaller and the rule becomes less interesting .

Piatetsky-Shapiro's principles apply to rule interestingness as long as the additional factors remain fixed. But in fact these additional factors do not remain fixed. These additional factors will probably vary a great deal across different rules, and this variation should be taken into account by the rule interestingness measure. There are 5 other factors related to the quality and interestingness of the rules [4]. These are:

- Disjunct size,
- The imbalance of the class distribution,
- Attribute cost,
- Misclassification cost,
- Asymmetry in classification rules.

II.II SMALL DISJUNCT PROBLEM

General idea is that small disjuncts are error-prone. Since they cover a small number of cases, it is possible that they cover mainly noise that can be generated in the rule extraction. At first glance, it seems that simply discarding small disjuncts can solve this problem. Unfortunately, prediction accuracy can be significantly reduced if the data-mining algorithm discards all small disjuncts [5]. This is a particularly serious problem in domains where the small disjuncts collectively match a large percentage of the number of cases belonging to a given class [6]. We face with the problem of choosing between two situations: Is a small disjunct a simple noise or a true exception of the data. In the later case the disjunct should be maintained, but in the former it is error prone and should be discarded. Unfortunately, it is very difficult to tell which is the case, by looking only at the data.

It is suggested that one remedy for the problem of small disjuncts could be to evaluate these disjuncts by using a bias different from the one used to evaluate large disjuncts [5]. This means that small disjuncts should be evaluated by a maximum-specificity bias, in contrast with the maximum-generality bias (favoring the discovery of more general rules) used by most data mining algorithms. Ting further investigated this approach, by using an instance-based learner (as far as we can go with the maximum-specificity bias) to evaluate small disjuncts [7].

II.III CLASS DISTRIBUTION IMBALANCE

A class distribution is unbalanced if rules belonging to one class are either much more frequent or much rarer than rules belonging to other classes.

A problem where two sample classes have almost the same frequency, is more difficult to solve than a problem where there is a great difference between the frequencies of the two classes. In the later case, it is easy to discover rules predicting the dominant class where in the former case it is difficult. The smaller the relative frequency of the minority class, the more difficult it is to discover rules predicting it and consequently, the more interesting are the rules predicting the minority class and the less interesting are the rules predicting the majority class.

II.IV ATTRIBUTE COSTS

In order to classify a new case with a given rule, it is necessary to use attribute costs. The algorithm must access the predicting attribute values of the new cases.

In some domains, different attributes might have very different costs to be accessed. For example consider a medical diagnosis problem. It is easy to determine the gender of the patient, but some health related attributes for example blood tests, can only be determined by performing a very costly examination.

Suppose that the antecedent of a discovered rule r_1 involves the result of an exam e_1 cost e.g. \$100 (X-Rays), while the antecedent of a discovered rule r_2 involves the result of another exam e_2 cost e.g. \$200 (tomography). All other things being equal, we would rather use rule r_1 for diagnosis. Hence r_1 becomes more interesting for us.

II.V MISCLASSIFICATION COST

Misclassification cost is the cost that we have to encounter, when the value of an attribute is predicted or determined incorrectly. For instance, the result of the medical examination of a patient is very important and misclassification cost of that attribute is very high. Suppose that a patient has cancer and the prediction is "Normal", and the result can be the death of the patient. But the cost of applying a chemotheraphy to a normal person is much lower with comparison to the death of the patient.

Using a rule interestingness measure which takes into account misclassification costs is not the only approach to deal with this problem. For instance, another approach consists of adjusting the relative proportions of each class in the data being mined.

II.VI ASYMMETRY IN CLASSIFICATION RULES

Classification is an asymmetric task with respect to the attributes in the database. In fact, we want to discover rules where the value of predicting attributes determines the value of the goal attribute, not vice versa. Hence, a rule interestingness measure should be asymmetric with respect to the rule antecedent and the rule consequent.

III INTERESTING RULE EXTRACTION

Piatetsky and Shapiro's (PS) measure is defined as $PS = |A\&B| - |A|\,|B|\,/N$. PS measure does take into account the imbalance of the class distribution by favoring the discovery of rules that predict the minority class. But other additional rule quality factors are not addressed. In [4] some extensions to PS measure have been proposed. In those extensions misclassification cost and attribute cost are taken into account as explained below, and other rule quality factors are rendered with the inclusion of those calculations.

III.I CALCULATING MISCLASSIFICATION COST

Different misclassifications have different costs. The cost of predicting that a patient does not have a disease, while he in reality does, is very high since he may die. The PS measure must be modified to take misclassification costs into account. In order to do that we should multiply PS formula by a new term called *InvMisClasCost* [4], which is the inverse of the sum of the expected misclassification costs as in the formula:

$$InvMisClasCost \;=\; \frac{1}{\sum_{j=1}^{k}\{Prob(j).Cost(i,j)\}}. \tag{1}$$

In Equation 1 *Prob (j)* is the probability that a case satisfied by the rule has true *class j*, *class i* is the class predicted by the rule, *Cost (i, j)* is the cost of misclassifying a case with true *class j* as *class i*, and k is the number of classes. Here, we have to explain *Cost (i, j)* and *Prob (j)*. *Cost (i, j)* as we explained before, is the cost of misclassifying a case with true *class j* as *class i*. We can see it clearly from Table 1.

When the prediction that the patient has cancer and in fact the patient is OK the 3^{rd} row is the cost that we have to pay for the medical treatment of the patient in case of the charge for the action of damage say 50000 dollars. And if we tell the patient that he has a heart problem the unnecessary treatment will cost 10000 dollars and suppose we have to pay again for our damaging action. In the 4^{th} row and 2^{nd} column we tell the patient that he is OK but in fact the patient is cancer. Then that number ($1,000,000) is the cost we have to pay to the patient's relatives for our action of damage after the patient dies. If the patient is cancer and we tell the patient that he is really cancer there is no cost for us. If the patient is in fact cancer and we say to him that he has a heart problem, the person will die and we will pay for both the patient's death and the unnecessary heart treatment he will have before he dies.

For *Prob (j)*, suppose we have 2 classes. Then a natural estimate for *Prob (j)* can be $|A\& \sim B| \,/\, |A|$ where $\sim B$ denotes the logical negation of the rule consequent. In order to take into account the interaction between misclassification cost and disjunct size, the reliability of this probability estimate can be improved by using Laplace correction. So $Prob(j) = (1 + |A\& \sim B|)/(2 + |A|)$.

If the number of classes is more than two, then the number 2 in the denominator is replaced with the number of classes.

Table 1. A sample misclassification table

Class J	Class I		
	OK	Cancer	Heart
OK	0	50000	10000
Cancer	1000000	0	1010000
Heart	1000000	1050000	0

III.II CALCULATING ATTRIBUTE COST

In our example application domain, attributes can represent several different kinds of predicting variables, including the patient's physical characteristics –gender, age, weight etc., and the results of the exams undergone by the patient –X-rays, blood tests, etc. Hence when we assign cost to each attribute, while costs we assign to the patient's physical characteristics are low, the cost assigned to the medical examinations will be relatively high.

In order to modify PS measure to take attribute costs into account, we multiply PS term with

$$AttUsef \;=\; \frac{k}{\sum_{i=1}^{k} Cost(A_i)} \tag{2}$$

as in [4], where k is the number of attributes occurring in the rule antecedent.

III.III RESULTANT INTERESTINGNESS CRITERION

The resultant interesting rule determination criterion is given in the Equation 3, which takes into account both Piatetsky–Shapiro's measures and the criteria given in [4].

$$\{|A\&B| - |A|\,|B|\,/N\} \times InvMisClasCost \times AttUsef. \tag{3}$$

As the value of Equation 3 increases the interestingness of the rule increases, which implies a direct relationship.

Table 2. The attributes of the data-set used in the experiments.

Classes	Clinical Features	Histopathological Features
C1: psoriasis	f1: erythema	f12: melanin incontinence
C2: seboreic dermatitis	f2: scaling	f13: eosinophils in the infiltrate
C3: lichen planus	f3: definite borders	f14: PNL infiltrate
C4: ptyriasis rosea	f4: itching	f15: fibrosis of the papillary dermis
C5: cronic dermatitis	f5: koebner phenomenon	f16: exocytosis
C6: pityriasis rubra pilaris	f6: polygonal papules	f17: acanthosis
	f7: follicular papules	f18: hyperkeratossis
	f8: oral mucosal involment	f19: parakeratosis
	f9: knee and elbow involment	f20: clubbing of the rete ridges
	f10: scalp involment	f21: elongation of the rete ridges
	f11: family history	f22: thinning of the suprapapillary epidermis
		f23: pongiform pustule
		f24: munro microabcess
		f25: focal hypergranulosis
		f26: disappearance of the granular layer
		f27: vacuolization and damage of basal layer
		f28: spongiosis
		f29: saw-tooth appearance of retes
		f30: follicular horn plug
		f31: perifollicular parakeratosis
		f32: inflammatory mononuclear infiltrate
		f33: band-like infiltrate
		f34: age

III.IV ALGORITHM

RULE DISCOVERY: We are in fact, concentrated on the analysis and presentation of the interesting rules. Therefore rule extraction from the data-set is done by VFI5 [8]. This program extracts rules using our 34 attributes. In the data-set there are 35 attributes which are shown in Table 2. The 35^{th} attribute is the actual disease of the patients.

Parameters for the rule discovery program are set in this way: Minimum probability of rules to be discovered is defined as 95%. Maximum number of cases and maximum number of conditions in a rule is set to 4. We wanted the rules to be produced as complex as possible, because many interesting rules can be discovered with this complexity. If we decrease the maximum number of conditions in the rule antecedent, i.e. 2, the number of discovered rules would be decreased but we would not be able to discover some of interesting rules, which have higher complexity. Since rule interestingness calculations do not include formulations, only if-then rules are checked for discovery.

ANALYSIS OF DISCOVERED RULES AND CALCULATING INTERESTINGNESS VALUES: After rule extraction is done, the discovered rules are exported to our program. With the help of queries exported, *IF-THEN RULES* query table is produced. This rule table provides the main source for our "Interesting Rule Extractor" program. Our algorithm is shown in Figure 1.

In this algorithm, at first necessary array elements in the main memory are created. This memory space will contain information about the coverage of each rule antecedent and consequent. After that, attribute costs and misclassification costs are read from their corresponding tables into the memory. Completing these, for each rule, its coverage and other parameters are calculated by making passes over data table. In each rule's coverage calculation, the corresponding field in the

```
Algorithm Find_Interesting_Rules
Begin
    Create arrays for coverage of A, B, A&B, A&~B
    Get Misclassification costs from the table
    Get Attribute costs from its corresponding table
    Select rule table
    Do While (Not EOF ())
        Find nonempty fields that contain conditions and condition numbers
        Parse each condition and find upper and lower limits of conditions
        Select data table
        Determine matching attributes with rules
        Do While (Not EOF ())
            Determine the coverage
        EndDo
        Calculate Misclassification Cost of the rule
        Calculate Attribute costs
        Calculate Interestingness of the rule
        Update interestingness field in the rule table
    EndDo
End.
```

Figure 1. Interesting rule extraction algorithm

rule table is filled with its interestingness value. After that the rule table is sorted in descending order according to the interestingness values.

VISUALIZATION OF INTERESTING RULES: The screen shot of the program is shown in Figure 2. In the interesting rule extractor program the user has the chance of changing attribute costs and misclassification costs at any time. The rows of misclassification cost represent our predicted classes with respect to the columns, which represent the true classes. This means that the first record of misclassification cost table corresponds to the cost of predicting classes 1-6 as class 1, the second row corresponds to the cost of predicting classes 1-6 as class 2, and so on. Since in our specific data-set we have six classes, there are six records. Also, attribute costs can be changed. There is only one record in the attribute cost table, corresponding to the handling cost of each attribute in the data-set.

In the middle of the form is the table of the original data-set, which contains real dermatology data. In the last part, rule table is given. Both rule table and original data table can be browsed to the right and left. The rule table contains attributes, which correspond to the attributes of the original data. To find what the rule is, the table should be browsed to the right while looking to a specific rule row. When a nonempty attribute is seen, we can understand that there is a condition. The value in the nonempty attribute shows the ranges of the attribute for that condition. Since the maximum number of conditions in each rule is four in our rule discovery program, it is able to see four attributes at maximum, which are nonempty. The rule table is in descending order according to the rule interestingness values.

After making necessary cost changes in the tables we can now recalculate the interestingness of each rule by pressing *Calculate* button in the bottom of the form, if needed. The program will now go over the steps of the algorithm explained above and sort our rule table according to the results of the calculations. After calculating the rule interestingness values, we can browse the rule table again and see which rules are more interesting in the perspective of the costs we have.

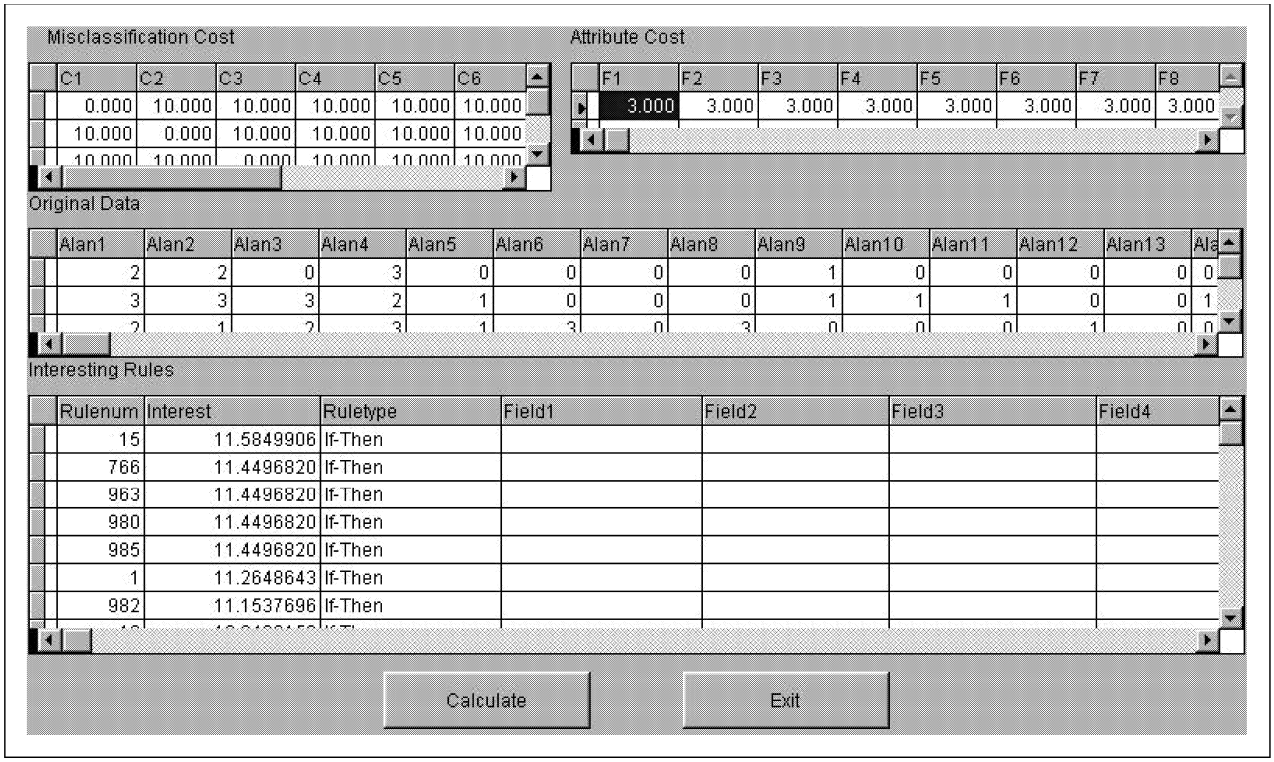

Figure 2. Snapshot of the interesting rule extractor program.

IV AN EXPERIMENT WITH THE INTERESTING RULE EXTRACTOR

Currently the data-set contains 366 records. The misclassification costs are shown in Table 3 and attribute costs in Table 4. These values are determined by dermatologists from Gazi University, at Ankara.

The values in the tables range from 0-5 and they are taken into account while determining the interestingness values as explained above. It should be kept in mind that in misclassification cost table, the values are not symmetric. For example, the misclassification cost of determining that a patient has disease 2 (predicted class-rows of table), instead of disease 6 (true class-columns of table) is 3 while the misclassification cost of determining that a patient has disease 6 which is in fact 2, is 2. The costs may be different with respect to the misclassification of the predicted attribute.

Table 3. Misclassification costs of the data-set

True Class	C_1	C_2	C_3	C_4	C_5	C_6
Predicted Class						
C_1	0	5	4	5	5	4
C_2	3	0	3	4	1	3
C_3	2	5	0	5	5	4
C_4	5	4	4	0	3	3
C_5	3	1	2	3	0	2
C_6	2	2	2	3	3	0

V RESULTS

The rules found as interesting were presented to the dermatologists at Gazi University. The dermatologists admitted that rules were correct and interesting. It should be kept in mind that correctness

Table 4. Attribute costs of the data-set

Clinical	Histopathological	
f1 = 1,5	f12 = 4,5	f23 = 4,5
f2 = 1,5	f13 = 4,5	f24 = 4,5
f3 = 1,5	f14 = 4,5	f25 = 4,5
f4 = 1,5	f15 = 4,5	f26 = 4,5
f5 = 1,5	f16 = 4,5	f27 = 4,5
f6 = 1,5	f17 = 4,5	f28 = 4,5
f7 = 1,5	f18 = 4,5	f29 = 4,5
f8 = 1,5	f19 = 4,5	f30 = 4,5
f9 = 1,5	f20 = 4,5	f31 = 4,5
f10= 1,5	f21 = 4,5	f32 = 4,5
f11= 1,5	f22 = 4,5	f33 = 4,5

Table 5. Top 10 interesting rules listed by Interesting Rule Extractor Program

Rule Number	Interestingness Value	Rule	Class
1	179,648816	if $f6 \geq 2$ and $f6 \leq 3$ then	f35 is in class 3
764	173,219672	if f6=0 and f7=0 and $f9 \geq 2$ and $f9 \leq 3$ then	f35 is in class 1
3	132,523133	if $f8 \geq 2$ and $f8 \leq 3$ then	f35 is in class 3
766	109,557741	if f6=0 and $f9 \geq 2$ and $f9 \leq 3$ and f31=0 then	f35 is in class 1
777	103,311803	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and f12=0 then	f35 is in class 1
782	103,931803	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and f25=0 then	f35 is in class 1
784	103,931803	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and f29=0 then	f35 is in class 1
783	101,174353	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and f27=0 then	f35 is in class 1
787	98,4539162	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and f33=0 then	f35 is in class 1
775	93,0253795	if f7=0 and $f9 \geq 2$ and $f9 \leq 3$ and $f10 \geq 2$ and $f10 \leq 3$ then	f35 is in class 1

is a subjective term for rule interestingness criteria. Especially in this kind of domains, correctness is not objective and may change from specialist to specialist (*specialist on dermatology*).

Our test platform is Intel Pentium III processor at 550 MHz with 64 MB of RAM. In this configuration, the algorithm finalized the work in 3 minutes and 55 seconds. This is a work of 366 records with 1881 rules. With that consideration and as the algorithm reveals, the running time of the algorithm is $\Omega(Rn)$, where R is the number of rules and n is the number of instances in the data-set.

The algorithm does not prune any rules, instead it sorts the rules according to their interestingness values. In that way the user has the ability to search for rules up to his or her interestingness threshold. Also with that ability the user also has the chance of comparing rules according to their interestingness.

Table 5 shows a collection of the most interesting 10 records from the calculated rule set with respect to the misclassification and attribute costs given in Tables 3 and 4. Note that, the interestingness values are just absolute value. Those values may show a great difference from calculation to calculation depending on your misclassification and attribute costs. So, comparing values from one calculation with other calculation values may mislead the user. Values must be compared in one calculation process. These rules are converted to a more literal representation, because in the program sorted rules are enlarging to the right and it is not possible to view them on one page since the attribute number is very large. In order to validate that, those rules are really interesting they should be verified by a large number of specialists, and also misclassification costs and attribute costs should be determined with the same group of specialists.

VI CONCLUSION AND FUTURE WORK

In this paper, an application, which extracts interesting rules from a target domain, is presented. As a matter of fact, our aim is not to develop a general application, which extracts interesting

rules, since interestingness is domain dependent and affected by many factors like, domain constraints, user expectations, etc. Hence, developing a general program is very complicated. Also if it is developed, it may not answer all interestingness criteria.

Our main goal is to show that interesting rule extraction can be done effectively and extensions presented in [3] and [4] are applicable and have a great effect on determining interestingness values, when domain specific effects are considered. We cannot overemphasize that a rule interestingness measure is a bias and so there is no universally best rule interestingness measure across all application domains. Each user must adapt a rule interestingness measure to the particular target domain they have.

Rule generation algorithms are not embedded into the rule learning program because we wanted to let the user have chance to select the most proper algorithm and generate the rules according to that algorithm. Other reason is that our purpose is not to present rule-mining algorithms but to focus on the extraction of interesting ones.

One open issue for the paper is that the developed program is strictly domain dependent. It may be expanded to take into consideration different domains. For future work we intend to work on our algorithm and its visualization methods, in order to make it faster and more presentable for the user. Some changes in the visualization process and data import sections can be done in order to improve it.

ACKNOWLEDGMENTS

We would like to thank Nilsel İlter from Gazi University, Department of Dermatology who provided the data-set and gave special interests on determining the attribute and misclassification costs for erythemato-squamous diseases.

References

1. U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: an overview," *Advances in Knowledge Discovery and Data Mining*, pp. 1–34, 1996.
2. B. Liu, W. Hsu, and S. Chen, "Using general impressions to analyze discovered classification rules," in *Proc. of 3rd Int. Conference in Knowledge Discovery and Data Mining*, 1997, pp. 31–36.
3. G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," *Knowledge Discovery in Databases*, p. 229, 1991.
4. A.A.Freitas, "On rule interestingness measures," *Knowledge Based Systems, Elsevier Science*, vol. 12, pp. 309–315, 1999.
5. R.C. Holte, L.E. Acker, and B.W. Porter, "Concept learning and the problem of small disjuncts," in *Proc. of Int. Joint Conf. AI (IJCAI-89)*, 1989, pp. 813–818.
6. A.P. Danyluk and F.J. Provost, "Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network," in *Proc. of . 10th Int. Conf. Machine Learning*, 1993, pp. 81–88.
7. K.M. Ting, "The problem of small disjuncts: its remedy in decision trees," in *Proc. of . 10th Canadian Conf. Artificial Intelligence*, 1994, pp. 91–97.
8. H.A.Güvenir, G.Demiröz and N.İlter, "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals," in *Artificial Intelligence in Medicine*, vol. 13, pp. 147–165, 1998.