# On Rule Interestingness Measures.

Alex A. Freitas
CEFET-PR (Federal Center of Technological Education), DAINF
Av. Sete de Setembro, 3165. Curitiba - PR, 80230-901, Brazil
Tel.: ++55 +41 322-4544 ext. 648
Fax: ++55 +41 224-5170 c/o Prof. Alex, DAINF
alex@dainf.cefetpr.br
http://www.dainf.cefetpr.br/~alex

## Abstract

This paper discusses several factors influencing the evaluation of the degree of interestingness of rules discovered by a data mining algorithm. The main goals of this paper are: (1) drawing attention to several factors related to rule interestingness that have been somewhat neglected in the literature; (2) showing some ways of modifying rule interestingness measures to take these factors into account; (3) introducing a new criterion to measure attribute surprisingness, as a factor influencing the interestingness of discovered rules.

Keywords: data mining, rule interestingness, rule surprisingness

## 1 Introduction

A crucial aspect of data mining is that the discovered knowledge should be somehow interesting, where the term interestingness arguably has to do with surprisingness (unexpectedness), usefulness and novelty [1].

Rule interestingness has both an objective (data-driven) and a subjective (user-driven) aspect. This paper focus on the objective aspect of rule interestingness. For a discussion about subjective aspects of rule interestingness, the reader is referred e.g. to [2]. It should be noted that, in practice, both objective and subjective approaches should be used to select interesting rules. For instance, objective approaches can be used as a kind of first filter to select potentially interesting rules, while subjective approaches can then be used as a final filter to select truly interesting rules.

This paper is organized as follows. Section 2 presents a review of several rule interestingness criteria. Section 3 presents a case study on how a popular rule interestingness measure can be extended to take into account several rule interestingness criteria in an integrated, combined fashion. Section 4 introduces a new criterion for rule interestingness measures. Finally, section 5 summarizes and concludes the paper.

# 2 A Review of Rule Interestingness Criteria

## 2.1 Rule Interestingness Principles

For the purposes of this paper, a classification rule is a knowledge representation of the form A => B, where A is a conjunction of predicting attribute values and B is the predicted class. When evaluating the quality of a rule, three common factors to be taken into account are the coverage, the completeness and the confidence factor of the rule, defined as follows. The coverage of the rule (i.e. the number of tuples satisfied by the rule antecedent) is given by |A|. The rule’s completeness (or proportion of tuples of the target class covered by the rule) is given by |A&B| / |B|. The rule’s confidence factor (or predictive accuracy) is given by |A&B| / |A|.

Piatetsky-Shapiro [3] has proposed three principles for rule interestingness (RI) measures, as follows.

1) RI = 0 if |A & B| = |A| |B| / N.

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.

The first principle says that the RI measure is zero if the antecedent and the consequent of the rule are statistically independent. The second and third principle have a more subtle interpretation. Note that Piatetsky-Shapiro was careful to state these principles in terms of *other parameters*, which is a phrase general enough to include any other parameter that we can think of. Let us assume for now that the rule parameters referred to by these principles are the terms |A|, |B|, and |A&B|, which are the terms explicitly used to state the principle. Note that this is an implicit assumption in most of the literature. However, we will revisit this assumption later in this section.

With the above assumption, principle 2 means that, for fixed |A| and fixed |B|, RI monotonically increases with |A&B|. In terms of the above 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 the higher these factors the more interesting the rule is.

Principle 3 means that: (1) for fixed |A| and fixed |A&B| (which implies a fixed coverage and a fixed confidence factor) RI monotonically decreases with |B| - i.e. the less complete, the less interesting the rule is; and (2) for fixed |B| and |A&B| (which implies a fixed rule completeness) RI monotonically decreases with |A| - i.e. the greater the coverage, the smaller the confidence factor, and the less interesting the rule is.

Major & Mangano [4] have proposed a fourth principle for RI measures (which does not follow from the first three principles), namely:

4) RI monotonically increases with |A| (rule coverage), given a fixed confidence factor greater than the baseline confidence factor (i.e. the prior probability of the class).

In passing, we mention that Kamber & Shinghal [5] have proposed a fifth principle for rule interestingness, but this principle is mainly oriented for characteristic rules, which are beyond the scope of this paper.

It should be noted that the above principles were designed mainly for considering the widely-used rule quality factors of coverage, completeness and confidence factor. Another widely-used rule quality factor is rule complexity. Although these factors are indeed important when evaluating the quality of a rule, they are by no means the only ones. In this paper we draw attention to five other factors related to rule quality and particularly to rule interestingness. These additional factors are discussed in the next subsections.

Note that, in theory, Piatetsky-Shapiro's principles still apply to rule interestingness measures considering these additional factors, as long as they remain fixed. (As mentioned before, the principles were carefully defined with the expression "fixed *other* parameters".) The problem is that, in practice, 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.

## 2.2 Disjunct Size

A rule set can be regarded as a disjunction of rules, so that a given rule can be regarded as a disjunct. The size of a disjunct (rule) is the number of tuples satisfied by the rule antecedent, i.e. |A|.

Thus, small disjuncts are rules whose number of covered tuples is small, according to some specified criterion (e.g. a fixed threshold, or a more flexible criterion). At first glance, it seems that small disjuncts are undesirable, and indeed most data mining algorithms have a bias favoring the discovery of large disjuncts.

Unfortunately, however, prediction accuracy can be significantly reduced if all small disjuncts are discarded by the data mining algorithm, as shown in [6]. This is a particularly serious problem in domains where the small disjuncts collectively match a large percentage of the number of tuples belonging to a given class [7]. The main problem is that a small disjunct can represent either a true exception occurring in the data or simply noise. In the former case the disjunct should be maintained, but in the latter case the disjunct is error prone and should be discarded. Unfortunately, however, it is very difficult to tell which is the case, given only the data.

Holte et al. [6] suggested that one remedy for the problem of small disjuncts was to evaluate these disjuncts by using a bias different from the one used to evaluate large disjuncts. Hence, they proposed that small disjuncts be evaluated by a maximum-specificity bias, in contrast with the maximum-generality bias (favoring the discovery of more general rules – i.e. larger disjuncts) used by most data mining algorithms. Ting [8] 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.

From a rule interestingness point of view, the lesson is that small disjuncts and large disjuncts should be evaluated in different ways – i.e. with different evaluation biases - by a rule interestingness measure.

## 2.3 The Imbalance of the Class Distribution

A class distribution is imbalanced if tuples belonging to one class are either much more frequent or much rarer than tuples belonging to other classes. To simplify our discussion, let us consider the common case of two-class problems.

Other things being equal, a problem where the two classes have the same relative frequency (or prior probabilities) is more difficult than a problem where there is a great difference between the relative frequencies of the two classes. In the latter case, it is relatively easy to discover rules predicting the majority class, but it is difficult to discover rules predicting the minority class. The smaller the relative frequency of the minority class, the more difficult it is to discover rules predicting it, and thus, intuitively, the more interesting are the rules predicting the minority class and the less interesting are the rules predicting the majority class. This point if often ignored by data mining algorithms.

Kononenko & Bratko [9] have proposed an information-theoretic measure for evaluating the performance of a classifier by taking into account the problem of imbalanced class distributions, and their measure has some interesting properties. However, their approach was designed to evaluate a classifier as a whole - mainly to compare the performance of different classifiers in the same domain or the performance of a classifier in different problem domains - rather than to compare the interestingness of different rules discovered by the same classifier, which is the focus of this paper.

Note that the problem of imbalanced class distributions interacts with other problems discussed in this paper. For instance, consider the interaction between the problem of imbalanced class distributions and the problem of small disjuncts. Let $r_1$ and $r_2$ be two small disjuncts (rules) of the same size (i.e. the same number of covered tuples), where $r_1$ predicts the minority class and $r_2$ predicts the majority class for a new tuple. Then $r_1$ tends to have a much smaller prediction accuracy than $r_2$ [10].

Finally, note that using a rule interestingness measure which takes into account the relative class frequencies is not the only approach to cope with the problem of imbalanced class distributions. For instance, another approach to address this problem consists of selectively removing tuples from the majority class, so that the class distribution becomes less imbalanced [11]. In this paper however, we are interested only in modifying the rule interestingness measure used by the algorithm, leaving the data being mined intact.

## 2.4 Attribute Costs

Most rule interestingness measures consider the rule antecedent as a whole, without paying attention to the individual attributes occurring in the rule antecedent. In this sense, these measures are coarse-grained. However, two rules with the same value of a coarse-grained rule interestingness measure can have very different degrees of interestingness for the user, depending on which attributes occur in the rule antecedent.

In this section we consider one situation where the notion of attribute interestingness is crucial and is related to the issue of attribute costs. In section 4 we will propose a new criterion to measure the interestingness of individual attributes occurring in a rule antecedent.

In order to classify a new tuple with a given rule, it is necessary to match the rule conditions against the tuple's predicting attributes (i.e. attributes other than the class one). Hence, the algorithm must access the values of the new tuple's predicting attributes. In some application domains, different attributes might have very different "costs" to be accessed. The typical example is medical diagnosis. For example, it is trivial to determine the gender of the patient, but some health-related attributes can only be determined by performing a very costly examination. In this case attribute costs must be taken into account when evaluating a rule. Continuing with our example, suppose that the antecedent ("if part") of a discovered rule $r_1$ involves the result of an exam $e_1$ costing, say, \$200, while the antecedent of a discovered rule $r_2$ involves instead the result of another exam $e_2$ costing, say, \$20. All other things (including prediction accuracy) being equal, we would rather use rule $r_2$ for diagnosis. In other words, the smaller the cost of the attributes occurring in the rule, the more interesting (the more useful, the less costly) the rule is. Some data mining algorithms that take into account attribute costs are described in [12], [13], [14].

## 2.5 Misclassification Costs

In some application domains, different misclassifications might have very different costs. For instance, in the domain of bank loans, the cost of erroneously denying  a loan to a good client (who is likely to pay it back) is usually considerably smaller than the cost of erroneously granting a loan to a bad client (who is unlikely to pay it back). In this case the data mining algorithm must be modified to take misclassification costs into account [15], [16], [17], [18]. This implies that the rule interestingness measure should take misclassification costs into account. We will revisit the issue of misclassification costs in section 3.2.1.

We must make here a comment similar to the one made in the section on imbalanced class distributions. Using a rule interestingness measure which takes into account misclassification costs is not the only approach to cope with this problem. For instance, another approach to address this problem consists of adjusting the relative proportions of each class in the data being mined. Once more in this paper, however, we are interested only in modifying the rule interestingness measure used by the algorithm, leaving the data being mined intact.

## 2.6 Asymmetry in Classification Rules

It should be noted that classification is an *asymmetric* task with respect to the attributes in the database. Indeed, we want to discover rules where the value of the predicting attributes determine the value of the goal attribute, not vice-versa. Hence, intuitively a rule interestingness measure should be asymmetric with respect to the rule antecedent and the rule consequent.

It is interesting to note that statistical measures of association, such as the popular $\chi^2$ (chi-squared) measure, which is widely used in data mining systems, were not designed for classification tasks. Rather, they were designed for measuring the association (or dependency) between two attributes in a *symmetric* way, i.e. none of the two rule terms (antecedent and consequent) being analyzed is given special treatment when computing the $\chi^2$ value.

We note in passing that an additional problem associated with the use of statistical significance tests in data mining, as pointed out by Glymour et al. [19], is that these tests were designed to evaluate a single hypothesis, whereas data mining algorithms typically have to evaluate many alternative hypothesis.

# 3 A Case Study on the Applicability of Additional Rule Interestingness Factors

The above subsections 2.2 through 2.6 have identified five factors that should be involved in measuring the interestingness of a rule, but that have often been somewhat ignored in the literature on rule interestingness. We now discuss how these factors can be applied to define a rule interestingness measure.

There are several rule interestingness measures proposed in the literature. As a case study, we will focus on one of the most popular ones, introduced by Piatetsky-Shapiro [3] as the simplest measure satisfying the three principles discussed in subsection 2.1. This measure, hereafter called PS (Piatetsky-Shapiro's) measure, is defined as:

$$PS = |A\&B| - |A||B|/N. \quad (1)$$

The remaining of this section is divided into two parts. Section 3.1 discusses how the PS measure addresses the additional rule interestingness factors discussed in subsections 2.2 through 2.6. Section 3.2 shows how this measure can be extended to better address some of those rule interestingness factors.

## 3.1 Analyzing the PS Rule Interestingness Measure

We now discuss how the PS measure, given by formula (1), addresses the rule quality factors of disjunct size, imbalance of the class distribution, attribute costs, misclassification costs and the asymmetry of classification rules.

*Disjunct size* - The PS measure takes into account the size of the disjunct, since formula (1) contains the term |A|. However, this measure treats small disjuncts and large disjuncts in the same way, with the same bias, which is undesirable, as discussed in section 2.2.

*Imbalance of the Class Distribution* - The PS measure takes into account the relative frequency (prior probability) of the class predicted by the rule, since formula (1) contains the term |B|. Other things being equal, the larger the value of |B|, the smaller the value of PS, so that the PS measure has the desirable property of favoring rules that predict the minority class.

*Attribute Costs* - The PS measure does not take into account attribute costs, neither any other measure of attribute interestingness. Actually, this measure considers the rule antecedent as a whole only, without paying attention to individual attributes of the rule antecedent.

*Misclassification Costs* - The PS measure does not take into account misclassification costs.

*Asymmetry of Classification Rules* - The PS measure is symmetric with respect to the rule antecedent and the rule consequent. We consider this an undesirable property of this measure, given the asymmetric nature of the classification task.

To summarize, out of the five factors influencing rule interestingness discussed in subsections 2.2 through 2.6, the PS measure takes into account only one of them (imbalance of the class distribution).

## 3.2 Extending the PS Rule Interestingness Measure

To render our case study more concrete, we will consider how to extend the PS rule interestingness measure in the context of a medical diagnosis application, where the goal is to predict whether or not the patient has a given fatal disease. We will make the realistic assumption that our application domain has two important characteristics, which will influence our design of an extended PS measure, namely varying misclassification costs and varying attribute costs. The next two subsections will discuss these two characteristics and how a rule interestingness measure can be extended to take them into account.

### *3.2.1 Varying Misclassification Costs*

Different misclassification have different costs. The cost of predicting that a patient does not have a disease, while (s)he in reality does, is very high, since it can lead to the death of the patient due to lack of proper treatment. On the other hand, the cost of predicting that a patient has a disease, while (s)he in reality does not, is relatively smaller – see also section 2.5. Hence, in our example application domain, the PS measure must be modified to take misclassification costs into account. A simple way of doing this is to multiply formula (1) by a new term called MisclasCost, defined as the inverse of the sum of the expected misclassification costs, as follows:

$$\text{MisclasCost} = 1 / \sum_{j=1}^{k} \text{Prob}(j)\text{Cost}(i,j), \quad (2)$$

where Prob(j) is the probability that a tuple 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 tuple with true class j as class i, and k is the number of classes.

Assuming a two class problem, a natural estimate for Prob(j) would be

$$\text{Prob}(j) = |A\&{\sim}B|/|A|, \quad (3)$$

where ~B denotes the logical negation of the rule consequent B. One problem with this estimate is that, if the rule covers few tuples, this estimate is not reliable. In other words, there is an interaction between the rule interestingness criteria of misclassification costs and disjunct size. Unfortunately, these criteria are usually considered independently from each other in the literature. In order to take into account the interaction between these two criteria, the reliability of the above probability estimate can be improved by using the Laplace correction [16], so that the estimate for Prob(j) in formula (3) would be given by

$$\text{Prob}(j) = (1 + |A\&{\sim}B|) / (2 + |A|). \quad (4)$$

(This correction can be easily generalized to an n-class problem by replacing the "2" in the denominator with n.) Note how the Laplace correction improves the reliability of a probability estimate for small disjuncts without significantly affecting this reliability for large disjuncts.

### *3.2.2 Varying Attribute Costs*

Different attributes have different costs of testing – see section 2.4. In our example application domain, attributes can represent several different kinds of predicting variables, including the patient's physical characteristics – e.g. gender, age, etc. – and the results of medical exams undergone by the patient – e.g. X-rays, blood tests, etc. Let us assume that each attribute has a well-defined cost, which represents the cost of determining the value of that attribute. Hence, attributes referring to the patient's physical characteristics have a minimum (virtually zero) cost to have their values determined, while attributes referring to the result of medical exams have much more significant costs to have their values determined.

Hence, in our example application domain, the PS measure must be modified to take attribute costs into account. A simple way of doing this is to multiply formula (1) by a new term called AttUsef (Attribute Usefulness), defined as the inverse of the sum of the costs of all the attributes occurring in the rule antecedent, that is:

$$\text{AttUsef} = 1 / \sum_{i=1}^{k} \text{Cost}(A_i), \quad (5)$$

where Cost($A_i$) is the cost of the i-th attribute occurring in the rule antecedent, and k is the number of attributes occurring in the rule antecedent.

Note that this formula has the side effect of penalizing "complex" rules, i.e. rules with many attributes in their antecedent. In some cases, however, the number of attributes in the rule is already being taking into account by another term of the rule interestingness measure, such as an explicit measure of rule complexity. In this case, to avoid that a rule be penalized twice for its high complexity, AttUsef can be simply defined as the inverse of the arithmetic average of the costs of all the attributes occurring in the rule antecedent, that is:

$$\text{AttUsef} = 1 / (\sum_{i=1}^{k} \text{Cost}(A_i) / k), \quad (6)$$

where Cost($A_i$) and k are as defined above.

To summarize, in our example application domain, the PS measure must be extended to take into account both misclassification costs and attributes costs, and a simple way of doing this is to multiply formula (1) by formulas (2) and (6). Notice that this extension also has the effect of rendering the PS measure asymmetric. It is easy to see that in other application domains the PS measure should be extended in other ways, depending on the particular characteristics of the application. Hence, a rule interestingness measures is a bias and, as any other bias, has a domain-dependent effectiveness [20], [21], [17]. The challenge is to define a rule interestingness measure that is the most suitable for the target application domain.

## 4 A New Criterion for Rule Interestingness Measures: Attribute Surprisingness

Sections 2.4 and 3.2.2 discussed attribute costs as a kind of rule interestingness factor. In the literature, this seems to be the only rule interestingness factor defined on a “fine-grain, predicting-attribute level” - i.e. directly based on individual attributes occurring in a rule’s antecedent - rather than being defined on a “coarse-grain” level, considering a rule antecedent as a whole. This section proposes a new rule interestingness criterion defined on the predicting-attribute level. Instead of focusing on attribute costs, which are related to rule usefulness, our new criterion focuses on the aspect of rule surprisingness. (Recall that rule interestingness involves several aspects, including both usefulness and surprisingness.)

Hence, we introduce a new term to measure rule surprisingness, called AttSurp (Attribute Surprisingness). In principle, any rule interestingness measure can be extended to take this term into account. For instance, the PS measure defined in formula (1) can be extended by multiplying that formula by the new term AttSurp. We propose that AttSurp be defined by an information-theoretic measure, based on the following idea.

First, we calculate the information gain of each attribute, defined as the class entropy minus the class entropy given the value of the predicting attribute. Attributes with high information gain are good predictors of class, when these attributes are considered individually, i.e. one at a time. However, from a rule interestingness point of view, it is likely that the user already knows what are the best predictors (individual attributes) for its application domain, and rules containing these attributes would tend to have a low degree of surprisingness (interestingness) for the user.

On the other hand, the user would tend to be more surprised if (s)he saw a rule containing attributes with low information gain. These attributes were probably considered as irrelevant by the users, and they are kind of irrelevant for classification when considered individually, one at a time. However, attribute interactions can render an individually-irrelevant attribute into a relevant one. This phenomenon is associated with surprisingness, and so with rule interestingness. Therefore, all other things (including prediction accuracy, coverage and completeness) being equal, we argue that rules whose antecedent contain attributes with low information gain are more interesting (more surprising) than rules whose antecedent contain attributes with high information gain. This idea can be expressed mathematically by defining the term AttSurp in the rule interestingness measure as:

$$\mathrm{AttSurp} = 1 / \left(\sum_{i=1}^{k} \mathrm{InfoGain}(A_i) / k\right), \quad (7)$$

where InfoGain($A_i$) is the information gain of the i-th attribute occurring in the rule antecedent and k is the number of attributes occurring in the rule antecedent.

# 5 Summary and Discussion

This paper has discussed several factors influencing the interestingness of a rule, including disjunct size, imbalance of class distributions, attribute interestingness, misclassification costs and the asymmetry of classification rules. These factors are often neglected by the literature on rule interestingness, which often focuses on factors such as the coverage, completeness and confidence factor of a rule.

As a case study, we focused on a popular rule interesting measure, defined by formula (1). We have shown that this measure takes into account only one of the five rule quality factors discussed in this paper, namely imbalanced class distributions. Then we discussed how this measure could be extended to take into account the other four factors. In particular, the extended rule interestingness measure has the form:

$$(|A\&B| - |A|\,|B| / N) * \mathrm{AttUsef} * \mathrm{MisclasCost}, \quad (8)$$

where the term AttUsef measures attribute usefulness - computed e.g. by formula (5) or (6) - and the term MisclasCost measures the misclassification cost - computed e.g. by formulas (2) and (4). Finally, the problem that formula (1) is symmetric, whereas classification rules should be asymmetric, was solved by adding the asymmetric terms AttUsef and MisclasCost to the extended formula (8).

The main goal of this paper was not to introduce yet another rule interestingness measure. Rather this paper had the main goals of: (1) drawing attention to several factors related to rule interestingness that have been somewhat neglected in the literature; (2) showing some ways of modifying rule interestingness measures to take these factors into account, which will hopefully inspire other researches to do the same; (3) introducing a new criterion to measure attribute surprisingness, as a factor influencing the interestingness of discovered rules. In particular, we believe that this new criterion is quite generic, and can be used in a large range of different application domains, so that it is a promising factor to take into account when designing a rule interestingness measure.

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 researcher or practitioner must adapt a rule interestingness measure (or invent a new one) to his/her particular target problem.

One limitation of this paper is that we have, implicitly, largely focused on how to measure the interestingness of different rules discovered by the same data mining algorithm, mining the same data. An open problem is how to extend our arguments for comparing the interestingness of different rules discovered by different data mining algorithms, or discovered from different data sets. Another limitation is that our discussion has not taken into account the interaction between rules in the induced rule set. In principle, however, the issue of rule interaction is somewhat orthogonal to the issue of individual rule interestingness, in the sense that the measure of rule interaction (typically a measure of rule overlapping) is often independent of the measure of individual rule interestingness. The reader interested in rule selection procedures taking into account rule interaction is referred to [22], [4], [23].

## References

[1] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth. From data mining to knowledge discovery: an overview. In: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. (Eds.) *Advances in Knowledge Discovery and Data Mining*, 1-34. AAAI/MIT Press, 1996.

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

classification rules. *Proc. 3rd Int. Conf. Knowledge Discovery & Data Mining*, 31-36. AAAI, 1997.

[3] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In: G. Piatetsky-Shapiro and W.J. Frawley. (Eds.) *Knowledge Discovery in Databases*, 229-248. AAAI, 1991.

[4]. J.A. Major and J.J. Mangano. Selecting among rules induced from a hurricane database. *Proc. AAAI-93 Workshop on Knowledge Discovery in Databases*, 28-44. July/93.

[5] M. Kamber & R. Shinghal. Evaluating the interestingness of characteristic rules. *Proc. 2nd Int. Conf. Knowledge Discovery & Data Mining,* 263-266. AAAI, 1996.

[6] R.C. Holte, L.E. Acker and B.W. Porter. Concept learning and the problem of small disjuncts. *Proc. Int. Joint Conf. AI (IJCAI-89),* 813-818.

[7] A.P. Danyluk & F.J. Provost. Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network. *Proc. 10th Int. Conf. Machine Learning*, 81-88, 1993.

[8] K.M. Ting. The problem of small disjuncts: its remedy in decision trees. *Proc. 10th Canadian Conf. Artificial Intelligence*, 91-97. 1994.

[9] Information-based evaluation criterion for classifier's performance. *Machine Learning 6*, 1991, 67-80.

[10] J.R. Quinlan. Improved estimates for the accuracy of small disjuncts. *Machine Learn.* 6(1), 93-98.

[11] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. *Proc. 14th Int. Conf. Machine Learning*, 179-186. Morgan Kaufmann, 1997.

[12] M. Nunez. The use of background knowledge in decision tree induction. *Mach. Learn.* 6, 231-250.

[13] M. Tan. Cost-sensitive learning of classification knowledge and its application in

robotics. *Machine Learning* 13, 1993, 7-33.

[14] P.D. Turney. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligent Research*, 2, Mar./95, 369-409.

[15] F. Provost & T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. *Proc. 3rd Int. Conf. Knowledge Discovery & Data Mining,* 43-48. AAAI, 1997.

[16] H. Roberts, M. Denby and K. Totton. Accounting for misclassification costs in decision tree classifiers. *Proc. Intelligent Data Analysis Conf. (IDA-95)*. 1995.

[17] D. Michie, D.J. Spiegelhalter and C.C. Taylor. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, 1994.

[18] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. *Classification and Regression Trees*. Pacific Groves, CA: Wadsworth, 1984.

[19]. C. Glymour, D. Madigan, D. Pregibon and P. Smyth. Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery* 1(1), 11-28. 1997.

[20] C. Schaffer. A conservation law for generalization performance. *Proc. 11th Int. Conf. Machine Learning*, 259-265. 1994.

[21] R.B. Rao, D. Gordon and W. Spears. For every generalization action, is there really an equal and opposite reaction? Analysis of the conservation law for generalization performance. *Proc. 12th Int. Conf. Mach. Learn.*, 471-479. 1995.

[22] F. Gebhardt. Choosing among competing generalizations. *Knowledge Acquisit.*, 3, 1991, 361-380.

[23]. J.A. Major and J.J. Mangano. Selecting among rules induced from a hurricane database. *J. Intel. Info. Systems* 4(1), Jan./95, 39-52.