Multi-Criteria Evaluation of Interesting Dependencies according to a Data Mining Approach

Dominique Francisci
I3S Laboratory,
University of Nice-Sophia Antipolis,
06903 Sophia Antipolis - Cedex France
francis@i3s.unice.fr

Martine Collard
I3S Laboratory,
University of Nice-Sophia Antipolis,
06903 Sophia Antipolis - Cedex France
mcollard@unice.fr

Abstract- This paper addresses the problem of the goodness of dependency rules extracted by mining data. Our approach is experimental and based on the idea that model quality has to be measured according to several criteria of quality such as accuracy, interestingness or domain-dependent criteria. Most works on model quality are focusing on one criterion at a time only and do not take into account multiple factors simultaneously. A few works combine different measures in weighted expressions. In order to combine multiple measures, we have first realized a comparative study which highlights the relative contribution of different factors and reveals trade-offs among some of them. This situation suggests looking in the rules search space for compromises rather than for best rules which may not exist. Thus we show that a multi-objective evolutionary approach is able to reveal interesting rules which are ignored by standard solutions.

1 Introduction

Data Mining may be defined as the discovery of unexpected relationships by analyzing such large volumes of data that automated processes are necessary. The extracted knowledge is expressed as a model or a pattern like sets of rules, neural networks or clusters for instance. As mentioned in [4] a model refers to a global structure that covers all the data while a pattern makes a summary or description of a locally restricted region of the data space. Numerous data mining applications are more concerned with the search of local patterns rather than global models. In this paper, we focus on rule-like patterns $A \rightarrow B$ where A and B are conjunctions of attribute-value terms *Att op a*, Att is an attribute, a is a value and $op \in \{=, <, >, \geq, \leq\}$. In a data mining process, it is quite frequent to search for rule based models since they are easily understandable by end users and have been found to be useful concepts for learning interpretable knowledge from data. One essential issue is to measure the interestingness of the dependency link between the premises A and the consequent B. Standard algorithms currently use basic measures for rule selection. But the specific goal of data mining has led to the definition of more specific indices which address different facets of rule goodness. In the following, we give an overview of standard algorithms and propose measures of interestingness.

1.1 Standard methods

Numerous algorithms have been proposed for rule induction from data in the machine learning literature. Tree induction or separate-and-conquer rule learning for instance are used for prediction in data mining too. They provide classification rules which right-hand side attributes are predefined and represent the class. These methods generally search for general prediction rules covering many data rather than specific rules. They aim to discover a set of rules that form an accurate classifier and this model is intended to apply to all the data. Thus the quality of an extracted model is evaluated according to its accuracy that is the ratio of correct classifications the entire set of rules makes on previously unseen examples.

As mentioned above, in frequent practical data mining applications, rules are rather considered individually outside the model they would form together. The goal is often to discover interesting patterns which are not only valid but novel, unobvious or surprising. These “good” patterns may not cover a large range of data. Criteria for goodness evaluation are measured on a rule individually rather than collectively.

One technique for generating a set of individually interesting and useful rules is to build a classification tree and then to evaluate each of the branches as individual rules according to specific targeted quality criteria. But some rules produced by standard classification make no sense to the user since they use biases and specific heuristics to generate the classifier. Thus this kind of technique may loose interesting rules as shown in [19].

On the other hand, so called association rules are among the most popular representation for local patterns in data mining. They originated in applications involving market-basket analysis where frequent relationships between products are expected to be discovered. In these rules the target is not predefined and the right-hand side of such a rule may be a conjunction of attribute-value terms. An association rule may be seen as a probabilistic statement about the co-occurrence of events which satisfy statistical constraints on the database like minimum $Support$ and minimum $Confidence$. The $Support$ is defined as $p(A \cap B)$ and $Confidence$ represents the conditional probability $p(B/A)$.

1.2 Dependency rules

Classification and association rules are two important techniques in data mining. Both of them are essential in practical applications. Besides this syntactical similarity, A. Freitas [8] highlights essential trade-offs between the two patterns. However while they refer to different concepts, one can observe they are merged for data mining applications. For instance, [2] proposes to integrate both kinds of rules in building classifiers which are more fitted to data mining objectives. This approach uses association rule mining techniques for classification tasks. The most popular association algorithm APriori [1] is adapted for this task. According to this technique, rules are selected if they satisfy minimum $Support$ and $Confidence$ which are not standard criteria for classification rules. Thus these rules predict a class-attribute value and they are considered individually. In [19] the system incorporates ideas from tree induction and rule learning. It is designed in order to search for classification rules for financial applications. The evaluation of rules comes in two flavors: number of errors on unseen data for the entire model, and $Confidence$ and $Support$ for individual rules. It is argued that in financial problems, it is unlikely that a model could be built that classifies all cases accurately. The system is designed to find models, even with few useful rules and with high accuracy. Rule goodness is determined on the basis of $Support$, $Accuracy$, $Entropy$ or some combination of such metrics on individual rules.

In this paper, we follow a similar approach: we study the interestingness and usefulness of rules which are considered individually. Furthermore, we look at rules as dependency links and we consider either a rule which right-hand side is a class-attribute or not. The framework is restricted to rules whose target is one attribute-value term only for convenience and simplicity. Our objective is to extend this study to more general rules. We do not make a distinction between classification rules or association ones since we are only interested in measuring the quality of the dependency link between premises and consequent of a rule.

1.3 Interestingness of dependencies

Simple criteria for rule selection like $Accuracy$ for entire model or $Support$ and $Confidence$ for individual rules are known to be insufficient for extracting useful and interesting information. For global classifier quality, and particularly for medical applications or information retrieval, $Sensitivity$ and $Specificity$ give more precision on the model quality. While usefulness and interestingness may have different meanings, one can distinguish objective and subjective approaches. Subjective criteria generally are based on a comparison of learned rules against an *a priori* knowledge on data.

Objective measures take their origin from the proposition of G. Piatetsky-Shapiro [14] who observed the weakness of the $Confidence$ factor and defined the $rule\ interest$. Numerous other measures have been proposed for evaluating the quality of the extracted information: the $Lift$ factor [12], the $JMeasure$ of Goodman and Smyth [10], the measure defined by Sebag and Schoenauer [16], the $Conviction$ [15]. It has been observed [4] that some of them are quite identical since they rank rules in the same manner. But they may provide a complementary approach since their definition was motivated in order to fill the gap.

1.4 A multi-criteria issue

Note that if standard algorithms are involved, the rule selection according to multiple criteria is made as post-processing the set of rules extracted first by the algorithm. This approach may result in undiscovered interesting rules. Few non-standard approaches have been proposed in order to apply a multiple criteria selection. For instance [11] and [5] use a genetic algorithm (GA) which fitness combines multiple factors in a weighted sum or a product. Note that these propositions may ignore rules which are good compromises according to different criteria.

In the context which we consider, i.e. the data mining and particularly the extraction of rules, few papers related to this metaheuristic in multi-objective optimization exist. For instance in [13], the authors suggest multi-criteria based metrics that can be used as comparators for an objective evaluation of data mining algorithms. They take into account all the available positive and negative features of these algorithms to construct a unique evaluation metric. Indeed, most existing studies only use the $Accuracy\ rate$ to compare algorithm performances. In [6], the author highlights the problem of feature selection which can be considered as a multi-objective problem since it generally evolves feature subset cardinality minimization and performance maximization. He uses the typical example of ROC analysis, where performance is assessed in terms of classifier $Specificity$ and $Sensitivity$. The author proposes an extension of the multi-objective evolutionary algorithm NPGA to handle $Specificity$ and $Sensitivity$ of classifiers. In the same way in [9], the authors use a GA to find the best subset of features that minimize both the error rate and the size of the tree discovered by a tree induction algorithm. Finally, we can quote [3] which use a GA in the marketing context, to select the customers likely to answer to a mailing request.

1.5 Our approach

In this paper, we argue that the opportunity of varying and combining different quality criteria is an essential advantage in a data mining process and we propose a evolutionary multi-objective method to address this problem. Our approach is experimental and based on the idea that rule quality is a multi-criteria issue as detailed above. We have chosen different interestingness measures and we experimentally have shown that they may be not correlated and even present trade-offs.

The paper is organized as follows. Section 2 presents a review of existing measures of interestingness. In Section 3, measures are compared by pair. Then Section 4 is devoted to the discovery of best compromises via a genetic algorithm which implements a multi-objective optimization. Section 5 summarizes and concludes the paper.

2 Interestingness Measures : positive correlations and trade-offs

In this section, we detail measures of quality proposed in the data mining literature. These measures may refer both to global model quality and individual rule quality. Even if they were defined for specific shapes like classifiers or associations, they take all sense to evaluate dependencies $A \rightarrow B$ as shown in [11] and [7].

2.1 Measures of interestingness

Classifier models expressed from tree induced from a training set are evaluated according to their $success\ rate$ on a test set i.e. the rate of truly classified examples in the test set. The $accuracy$ of a rule $A \rightarrow B$ is defined in terms of probability by $p(A \cap B) + p(\neg A \cap \neg B)$. The $sensitivity$ and $specificity$ criteria were defined for a classifier in order to give a more precise idea of its performance. They specially take sense when class distribution is unbalanced. For a rule $A \rightarrow B$, $sensitivity$ becomes $p(A \cap B)/p(B)$ or $p(A/B)$ and evaluates the coverage of B by A while the $specificity$ becomes $p(\neg A \cap /\neg B)/p(\neg B)$ or $p(\neg A/\neg B)$ and evaluates the coverage of $\neg B$ by $\neg A$. The choice of a quality measure depends on the specific facet of interestingness we want to favor. For instance, in a classification task for a medical diagnostic test, the main objective is to reduce the error which consists in predicting a patient in class "healthy". In this case, experts from the domain may consider it is more essential to have best results in the classification of examples of class "patient" even if examples from class "healthy" are classified as "patient". This means to optimize $sensitivity$ rather than $specificity$. In other domains, one may want to optimize both $sensitivity$ and $specificity$ simultaneously.

For association rules, maximizing $support$ allows to select frequent patterns, but $confidence$ has proved its limits in filtering them since it favors numerous rules which are most often irrelevant. Indeed the weakness of the $confidence$ factor $p(B/A)$ is to ignore $p(B)$. Most interestingness measures generally compare the $a\ priori$ probability $p(B)$ and the $a\ posteriori$ probability $p(B/A)$. For instance, the $lift$ measure [12] defined as $p(B/A)/p(B)$ and the $Rule\ Interest$ measure from G. Piatetsky-Shapiro defined as $|A|\,(p(B/A) - p(B))$ measures the departure from independence between A and B. As remarked by [15], they measure the co-occurrence of A and B since they are symmetric. The $conviction$ [15] defined by $p(A)p(\neg B)/p(A \cap \neg B)$ is not symmetric and is related to the logical implication $A \rightarrow B$ which means $\neg A \vee B$. M. Sebag and M. Schoenauer [16] defined the measure $p(A \cap B)/p(A \cap \neg B)$ which measures the ratio between positive and negative examples. The $JMeasure$ [10] defined as $p(A \cap B)log(p(A \cap B)/p(A)p(B)) + p(A \cap \neg B)log(p(A \cap \neg B)/p(A)p(\neg B))$ by combining a factor of generality and a factor of *goodness-of-fit* quantifies the information content of the rule. The rule quality is not reduced to one criterion only. Furthermore, the goodness cannot be specified in an absolute way, it depends on specific goals of the search process. For instance one can search for large rules which cover a whole class. On the contrary it may be useful to find small rules which quality involves rarity and precision. Thus the questions we address here are : how to combine multiple criteria simultaneously ? Is there equivalence among measures ? What is the best solution for finding the best rules according to multi-criteria ?

2.2 Correlations and Trade-offs

Some measures are known to be strongly correlated particularly if the search space is reduced by fixing p(B). This is the case when we search for classification rules on one given class. For instance $rule\ Interest$ and $support$ rank rules in the same order if p(B) is constant. The three measures $lift$, $conviction$ and $sebag$ have the same property. For a given class, $sensitivity$ and $specificity$ represent complementary indices of coverage. One can recall that these two factors are plotted in ROC curves in order to evaluate the performances of a classifier against positive and false negative examples.

Since trade-offs may exist, it is quite obvious that standard techniques which are not designed for multi-objective optimization may ignore good compromises. In order to make an experimental study, we have identified several cases where trade-offs are apparent. To highlight correlations among couples of measures, we have plotted sets of points where each point maps a rule measured according to two measures. We have studied randomly generated rules which are expected to represent the search space. Resulting graphs below show three situations of trade-off. Thus a multi-objective optimization method has been implemented on these three cases to elicit best compromises. Thereafter the presentation focuses on these cases. Case 1 is related to criteria $Sensibility$ (Se) and $Specificity$ (Sp) which characterize the precision and the non-coverage of negative examples and are good indicators of rule quality in diagnostic test and document retrieval. In case 2, we study two common factors which are $Support$ $(Supp)$ and $Conviction$ $(Conv)$. Finally case 3 shows the relative behavior of the $Rule\ Interest$ (RI) and the $Sebag$ coefficient.

Experiences have been built on data files from the UCI repository [18]. Randomly generated rules are useful for estimating the density of the rule search space according to values of criteria but they obviously may not give a complete vision of it. For each data file up to 10^5 rules were generated.

In this paper, for simplicity reasons, we consider rules $A \rightarrow B$ where B is one attribute-value term only.

On Figure 1, we can observe results for the couple $(Sensitivity, Specificity)$ on the $Vote$ data set. This data set contains 435 descriptions of politician candidates according to 17 categorical attributes. These data were used for classification. The class attribute $politic\ class$ takes value "$democrat''$ or "$republican''$. By observing Figure 1 it seems that there is not any optimal rule : not any rule is plotted in the upper right corner. This graph suggests an apparent trade-off between the two measures on this dataset. A solution to obtain best compromises may be to optimize the product on the rule set discovered by a standard algorithm.

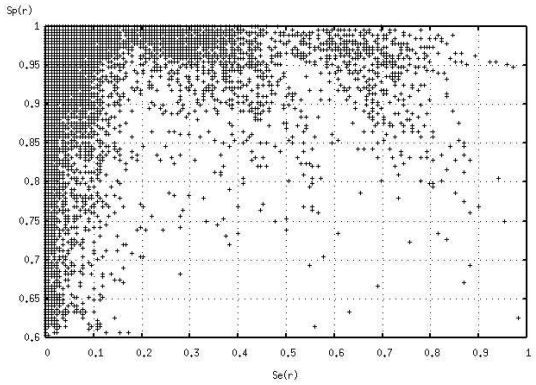

Figure 1: Random rules evaluated according to $Sensitivity$ and $Specificity$

$Rule\ Interest$ as a symmetric criterion measuring the departure from independence between A and B and $Sebag$ [16] which is sensitive to the negative examples, present an apparent tradeoff too like we can see in Figure **??**. The graphical representation highlights the correlation between the two measures when RI is negative. We can observe that $Sebag$ is closed to zero in that case ; indeed, in this situation $p(A \cap B)$ tends to be small and thus it tends to be less than $p(A \cap \neg B)$.
The $Conviction$ index introduced in [15] to replace the $confidence$ is used in an extension of the two-steps $APriori$ algorithm as follows : first the frequent itemsets of minimum $support$ value are selected and then itemsets with high $conviction$ value are kept. However, we can see on Figure 2, from a particular value of $support$, for instance 0.3, some solution would be discovered because of their relative high $support$ value, but some others solutions with high $conviction$ value would be ignored by an $APriori - like$ algorithm. For instance, so called $small\ disjuncts$ which are rules covering a small number of examples, but may be good according to their $conviction$ value would not be discovered. This is a typical issue to address with a multi-criteria optimization method.

The $Rule\ Interest$ [14] as a symmetric criterion measuring the departure from independence between A and B and the $Sebag$ coefficient [16] which is sensitive to the negative examples, presents the same aspect like we can see in Figure 3. Indeed, for a particular value of $Sebag$, there

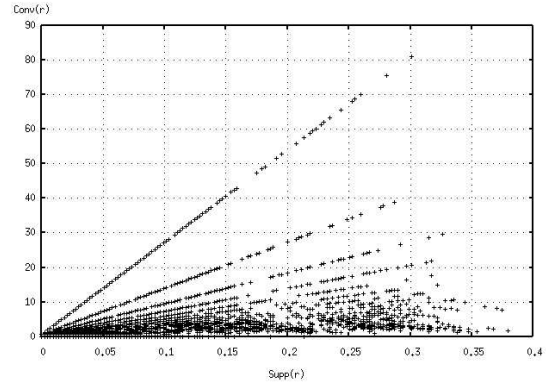

Figure 2: Random rules evaluated according to $Support$ and $Conviction$

exist several good solutions according to $Rule\ Interest$.

3 Multi-Criteria Selection Rule

This section presents the results we have obtained with an evolutionary optimization method. We have applied the NSGA (Non dominated Sorting Genetic Algorithm)[17] which allows to get a good diversity into solutions.

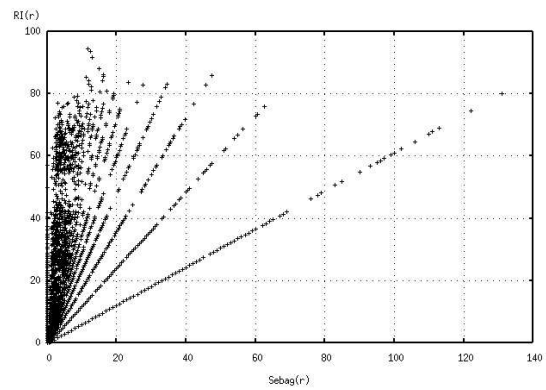

Figure 3: Random rules evaluated according to $Sebag$ factor and $Rule\ Interest$

3.1 Multiobjective optimization

The *multiobjective* optimization or *Pareto* optimization can be defined mathematically as follows :
each solution X_i is associated with an evaluation vector $F = (F_1(X_i), ..., F_N(X_i))$ where N is the number of objectives. One solution X_1 is said to $dominate$ another solution X_2 if $\forall j : F_j(X_1) \geq F_j(X_2)$ and $\exists k : F_k(X_1) > F_k(X_2)$, where F_j and F_k are respectively the j-th and k-th objectives, $i, j \in \{1, ..., N\}$. Neither solution dominates the other if $\exists m_1, m_2 : F_{m_1}(X_1) > F_{m_1}(X_2), F_{m_2}(X_2) > F_{m_2}(X_1)$. The *Pareto front* is defined as the set of non dominated solutions. First, the goal is to approximate as best as possible the Pareto front.
The objective functions represent in our case the measures

of extracted rules and the variables X_i represent the rules. Genetic algorithms (GA) are related to the mechanisms of natural evolution, that mainly is to say selection, reproduction and mutation. They describe the way individuals in a population evolve in response to their environment over successive generations and are selected according to the principle of the survival of the most fitted.

GAs belong to the metaheuristic family built to solve difficult optimization problems. From the data mining point of view, they have mainly the advantage to allow evaluating rules globally while other methods like decision tree induction are evaluating one term at a time. Furthermore, one of the main difference of the GA compared to standard optimization algorithms, is that these algorithms use a population of solutions instead of a single solution. If a problem has several optima, a GA is able to capture these different local optima and this is done during a single run.

In the implementation we present in this paper, each individual represents either the antecedent of a rule or the entire rule. In the first case the dependency rule is simple : its *conclusion* part is fixed at the beginning of the execution and represents a goal attribute value. Thus, in that case to discover several rules predicting different goal attribute values, it is necessary to run the GA several times, one for each value. In the second case, the GA is searching for solutions in the whole search space of dependency rules.

The algorithm we present here is based on the NSGA (Non dominated Sorting Genetic Algorithm) algorithm previously described in [17] . It uses the Pareto dominance to determine the fitness of an individual as follows : each individual has a rank which represents the number of individuals in the current population which dominate it. This rank is defined as follows for an individual i : $rank(i) = 1 + p(i)$ where $p(i)$ is the number of individuals which dominate the i individual. Non dominated individuals are assigned the rank 1 and dominated individuals have a greater rank. The individuals of same rank belong to a same category which is assigned to a virtual fitness value equals to $1/P$ with P the Pareto rank of this category. To obtain a diversity of individuals in this category (or ecologic niche), the fitness value of individuals is modified in $fitness/m_i$ where the bias m_i is equal to $\sum_{j=1}^{K} P(d(i,j)$ and $P(d(i,j)) = 1 - \frac{d(i,j)}{disInf}$ if $d(i,j) < disInf$ and 0 otherwise. K is the number of individuals in the considered category and $d(i,j)$ is the Euclidian distance (based on the two measures considered) between individuals i and j. The distance $disInf$ is the distance of influence and defines the radius of the ecologic niche.

The GA aims at finding the best approximation of the Pareto set of solutions. Ideally, this set must be contained in the final population. Nevertheless, this result is not guaranteed. Indeed, the stochastic nature of the GA via mutation operator may lead non dominated individuals to disappear. On the other hand, the extent of the Pareto front is not necessary known a priori and, therefore, the population dimension in consequence cannot be determined precisely. These two phenomena suggest archiving non dominated solutions found during the GA execution. When a non dominated solution is met by the GA due to a crossover or a mutation, it is stored into the archive. Finally solutions presented to the user are those which are present in the archive, they give an approximation of the Pareto front but they may not be solutions contained in the final population as usually.

3.2 Rule Extraction with the multi-objective GA

This section presents our approach towards the individuals encoding, the genetic operators and the fitness function. It shows the results we have obtained by the implementation of this method and its application on several databases.

3.2.1 Individuals representation

For rule extraction we have chosen the Michigan approach of GA : an individual codes a single rule. An individual genome represents the conjunction of attribute-value terms corresponding to the rule antecedent. The coding is a position coding consisting in a sequence of genes ranked in the same order than corresponding attributes in the data set. Each condition is coded by a gene and consists in an attribute-value *Att op a* where Att is an attribute, op is one of the operators $=$, $<$, $>$, $\leq$ or $\geq$ and a is a value of Att. Each gene is composed with a Boolean field which indicates if the gene is active or not, that is to say if the condition is present in the rule. So, while individuals have the same genotype length, associated rules are length-variable. The algorithm can handle integer, real or categorical variables. For a categorical attribute, the only operator is $=$ and for integers and reals available operators are $=$, $<$, $>$, $\leq$ and $\geq$.

3.2.2 Genetic operators

The crossover we used consists first in randomly choosing a site. Then it exchanges the genetic material located after this site in both individuals. The site is chosen to avoid the generation of empty individuals. The mutation operator modifies the rule length. Either, it specializes a rule by the adjunction of a condition if at least one gene is not active, or it generalizes it by deleting a condition. The adjunction of a condition is made by randomly generating an operator and a value belonging for the mutated gene. The selection we used is the tournament selection with two individuals.

3.3 Experimental results

We have chosen to apply the multi-objective GA on $(Sensibility, Specificity)$, $(Sebag, Rule\ Interest)$ and $(Support, Conviction)$ couples. The GA has been tested on the $Vote$ dataset presented in section 2.2. The GA parameters have been widely tested according to data sets characteristics. For instance, on the $Vote$ dataset, we have retained : 200 individuals in the population, 100 generations, and a distance of influence equal to 0.05. The crossover probability is 0.8 and the mutation probability is 0.05.

In the following graphs are representing a rule space. Each graph contains three types of information: we have generated 10^5 random rules for a representation of the search space and we have plotted rules discovered by the multi-objective GA and rules discovered by a tree induction algorithm. The following parameters have been used in the GA for each graph : population size : 200 individuals, number of generations : 50, crossover rate : 0.8, mutation rate : 0.0625 and distance of influence : 0.05.

The figure 4 shows resulting rules obtained with the multi-objective GA. In this first example, we can observe discovered rules on the Pareto frontier for the $Sensitivity$ and $Specificity$ measures. For instance, the following rule with $(Se = 0.87, Sp = 0.96)$ is on this frontier:

IF (physician-fee-freeze=y) **AND** (el-salvador-aid=y) **AND** (duty-free-exports=n) **THEN** Class=republican

We can observe on this graph too that a tree induction algorithm is not able to find the same compromises since it only optimize one criterion.
Figures 5 and 6 present respectively the results obtained with the multi-objective GA for the $(Sebagfactor, Rule\ Interest)$ and $(Support, Conviction)$.

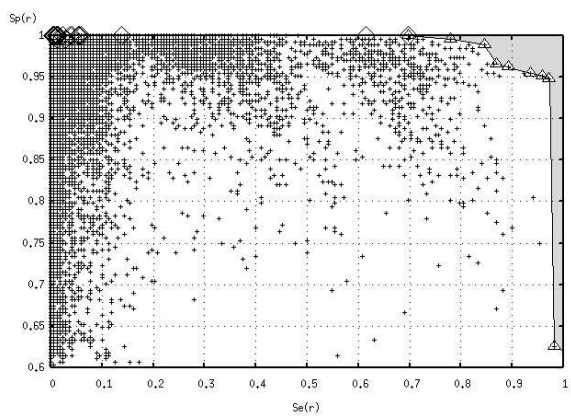

Figure 4: $Sensitivity$ against $Specificity$. Cross : random method - Triangles : GA - Rhombus : Tree induction

Most rules obtained by the multi-objective GA are not discovered by the tree induction algorithm. Here are specific rules among these ones. The first one with $(Se = 0.89286, Sp = 0.96255)$ generalizes several rules resulting from tree induction. The second one whose coordinates are $(Se = 0.95833, Sp = 0.95131)$ does not share any common term with tree induction resulting rules.

IF (physician-fee-freeze=y) **AND** (duty-free-exports=n) **THEN** class=republican

IF (physician-fee-freeze=y) **AND** (crime=y) **THEN** class=republican

The rules found by tree induction have a $Specificity$ index equal to 1. They have also a very low $Sensitivity$ while rules found with the GA are compromises. For instance, if we consider the context of the medical domain and particularly the diagnostic of diseases, the rules found by tree induction are not the best..

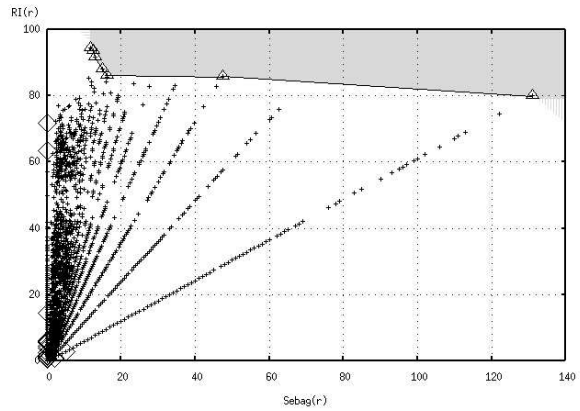

Figure 5: $Sebagfactor$ against $Rule\ Interest$. Cross : random method - Triangles : GA - Rhombus : Tree induction

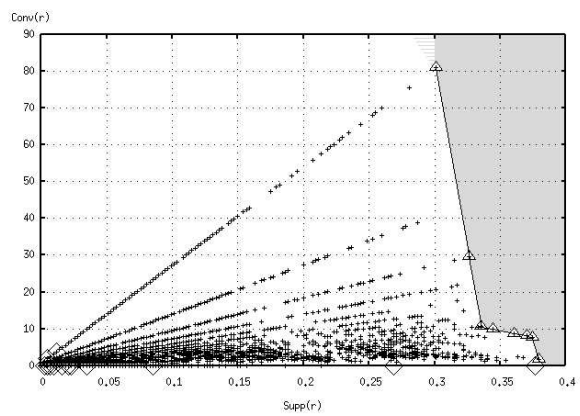

Figure 6: $Support$ against $Conviction$. Cross : random method - Triangles : GA - Rhombus : Tree induction

4 Conclusion

In this paper, we have proposed a multi-criteria approach for extracting rules according to different criteria of goodness. Indeed, in order to answer to the main objective of data mining, it becomes essential to be able to select useful information. Usefulness means not only interestingness, surprisingness but precision, comprehensibility and domain-dependent qualities too. Even only one of these qualities, interestingness, may be appreciated differently according final objectives. Numerous criteria and measures have been defined and implicitly considered as sufficient. Some of them are known to be equivalent and others

are complementary. The comparative study we have led on different measures has shown several phenomena of tradeoffs between them. According to these observations we have addressed the problem of selecting interesting rules as a multi-criteria optimization problem. A multi-objective evolutionary algorithm has been chosen since this kind of stochastic methods is able to explore large search spaces and allows defining easily different evaluation functions. Thus this approach emphasizes the idea that interestingness cannot be defined in an absolute way. It allows to easily parameterize the rule search process and to make varying the parameter values according to different and multiple goals. Next steps on this work will focus on combining this preliminary multi-objective approach which provides a selection of objectively interesting patterns with subjective interestingness filters.

Bibliography

[1] R. Agrawal and A. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB 1994*, pages 487–499, 1994.

[2] W. Hsu B. Liu and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-1998), Plenary Presentation)*, New York, 1998.

[3] S. Bhattacharyya. Evolutionary algorithms in data mining : Multi-objective performance modeling for direct marketing. In *Proc. of the 6th International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, Boston, August 2000. AAAI Press.

[4] H. Mannila D. J. Hand and P. Smyth. *Principles of data mining*. Adaptive Computation and Machine Learning. J. Wolfskill, 5 Cambridge Center, 4th floor, Cambridge, MA 02142, 2001.

[5] H.S. Lopes D.L.A. Araujo and A.A. Freitas. A parallel genetic algorithm for rule discovery in large databases. In *Proc. 1999 IEEE Systems, Man and Cybernetics Conf., v. III*, pages 940–945, Tokyo, October 1999.

[6] C. Emmanouilidis. Evolutionary multi-objective feature selection and roc analysis with application to industrial machinery fault diagnosis. In *Eurogen 2001*, 2001.

[7] I. W. Flockart and N. J. Radcliffe. Ga-miner : Parallel data mining with hierarchical genetic algorithms, final report. In *EPCC AIKMS GA-Miner - Report 1.0*, University of Edimburgh, November 1995.

[8] A.A. Freitas. A multi-criteria approach for the evaluation of rule interestingness. data mining. In *Proc. International Conference*, pages 7–20, Rio de Janeiro, Brazil, September 1998. WIT Press.

[9] A. A. Freitas G. L. Pappa and C. A. A. Kaestner. A multiobjective genetic algorithm for attribute selection. In A. Lofti J. Garibaldi and R. John, editors, *Proc. of the 4th International Conference on Recent Advances in Soft Computing (RASC-2002)*, pages 116–121. Nottingham Trent University, December 2002.

[10] R. M. Goodman and P. Smyth. Rule induction using information theory. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases (KDD-1991)*. MIT Press, 1991.

[11] N. J.Radcliffe and P. D. Surry. Co-operation through hierarchical competition in genetic data mining. In *EPCC-TR94-09*, 1994.

[12] International Business Machines. *IBM intelligent Miner, User's guide*. IBM, 1996.

[13] G. Nakhaeizadeh. Development of multi-criteria metrics for evaluation of data mining algorithms. In D. Pregibon D. Heckerman, H. Mannila and R. Uthurusamy, editors, *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-1997)*, pages 37–42, Newport Beach, California, August 1997. AAAI Press.

[14] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and editors W. J. Frawley, editors, *Knowledge Discovery in Databases*. MIT Press, 1991.

[15] J.D. Ullman S. Brin, R. Motwani and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD-1997)*, Tucson, Arizona, USA, May 1997.

[16] M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of the European Knowledge Acquisition Workshop EKAW-1988)*, 1988.

[17] N. Srinivas and K. Deb. Multiobjective optimization using non-dominated sorting in genetic algorithms. Technical report, Department of Technical Engineering, Indian Institute of Technology, Kanput, Indian, 1993.

[18] Irvine University. Uci machine learning. ftp.ics.uci.edu/pub/machine-learning-databases.

[19] D. Chou V. Dhar and F. Provost. Discovering interesting patterns for investment decision making with glower - a genetic learner overlaid with entropy reduction. *Data Mining and Knowledge Discovery Journal*, 4(4):251–280, 2000.