

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

A SCALAR EVOLUTIONNARY APPROACH TO RULE EXTRACTION

Dominique FRANCISCI, Laurent BRISSON, Martine COLLARD

Projet MECOSI

Rapport de recherche
ISRN I3S/RR-2003-12-FR

Mai 2003

RÉSUMÉ :

Cet article s'intéresse à l'utilité des règles découvertes en fouille de données. Notre approche est expérimentale et basée sur la remarque suivante : la qualité des connaissances extraites dans un processus de fouille de données doit être évaluée selon plusieurs critères comme la précision, la nouveauté, la compréhensibilité ; cependant les travaux sur ce thème ne prennent pas en compte l'intégration de plusieurs facteurs. Nous présentons quelques résultats de l'étude comparative que nous avons menée sur le comportement relatif de différentes mesures de qualité. Nous traçons des nuages de points représentant des règles caractérisées par la donnée de deux mesures. Nous observons certains phénomènes et en particulier des antagonismes entre mesures. Cette situation suggère de parcourir l'espace de recherche des règles pour découvrir des compromis plutôt que des règles optimales. Nous étudions différents cas et montrons l'intérêt d'une approche scalaire pour découvrir les règles optimisant plusieurs critères. Pour implémenter cette solution, nous avons choisi un algorithme génétique apte à parcourir des espaces de recherche très important et permettant de combiner aisément plusieurs critères par l'intermédiaire de sa fonction de fitness.

MOTS CLÉS :

fouille de données, extraction de règles, critères de qualité, approche scalaire, algorithme génétique

ABSTRACT:

In this paper, we are interested in the interestingness of rules extracted from data. Our approach is experimental and based on the following remark : the quality of discovered rules in data mining process has to be measured according several criteria such as accuracy, interestingness and comprehensibility, but most work on rule interestingness do not take into account the integration of multiple factors. We present a comparative study which highlights the relative behaviour of different quality measures. We plot sets of rules according a pair of criteria. We obtain graphical representations which show several antagonisms among quality measures. This situation suggests to look in the rule search space for compromise rather than for best rules. We study different solutions and show the advantage of a scalar method for multi-criteria optimisation. The solution is implemented via a genetic algorithm. We have run the algorithm several times in order to propose a set of compromise to the user.

KEY WORDS :

data mining, rules extraction, quality criteria, scalar approach, genetic algorithm

A Scalar Evolutionary Approach to Rule Extraction

Dominique Francisci, Laurent Brisson, Martine Collard

I3S Laboratory - University of Nice-Sophia Antipolis
Les Algorithmes - Bât. Euclide B
2000 route des Lucioles, B.P. 121
06903 Sophia Antipolis Cedex - FRANCE
{francis@i3s.unice.fr, brisson@essi.fr, mcollard@unice.fr}

Abstract This paper addresses the problem of the goodness of models extracted by mining data. Our approach is experimental and based on the idea that model quality has to be measured according several criteria such as accuracy, interestingness or domain-dependent criteria. Most works on model quality are focusing on interestingness only and do not take into account multiple factors simultaneously. In order to combine multiple measures, we have first realized a comparative study which highlights the relative contribution of different quality measures and reveal antagonisms among some of them. This situation suggests to look in the rule search space for compromises rather than for best rules which may not exist. Existing solutions are studied and a scalar method for multi-criteria optimisation is proposed. This solution is implemented via an evolutionary algorithm.

1 Motivations

Data Mining may be defined as the discovery of knowledge which is hidden in such large volumes of data that automated processes are necessary. In standard mining algorithms, criteria for model selection are mostly based on accuracy. But this measure is known to be insufficient for extracting useful and interesting informations. Numerous measures of interestingness have been proposed : the *RI* measure of Piatestsky-Shapiro [4], the *lift* factor [3], the *J-Measure* of Goodman and Smith [7], the measure defined by Sebag and Schoenauer [6]. In this paper, we study rule goodness as a multi-criteria problem. Our approach is experimental and based on the idea that rule quality has to represent not only interestingness but other criteria like accuracy, comprehensibility, and even domain-dependent factors. But researches on model quality generally do not propose to integrate multiple criteria in the model selection process.

The paper is organized as follows. Section 2 presents a review of existing measures of interestingness for rules. In Section 3, measures are compared. Then Section 4 is devoted to the discovery of best compromises via a genetic algorithm which implements a scalar optimisation. Section 5 summarizes and concludes the paper.

2 What does goodness mean about rules?

In this paper, we consider models as sets of rules $A \rightarrow C$ where A and C are conjunctions of itemsets. In this context, a rule may be a *classification* rule or an *association* rule. We do not make any distinction between them since we are interested in the causal link between A and C only. *C4.5* [5] and *APriori* [1] are the most popular algorithms for respectively extracting these rules and we use them as references. These algorithms base rule selection on accuracy measures like *success rate* for *C4.5*, *support* and *confidence* for *APriori*. *C4.5* is a tree induction method and classifier models expressed from tree induced from a training set are evaluated according their success rate on a test set (i.e the rate of truly classified examples in the test set).

Other measures have been defined like *sensitivity* and *specificity* : for a rule $A \rightarrow C$, while the success rate is equal to $p(A \cap C) + p(\bar{A} \cap \bar{C})$, *sensitivity* Se is equal to $p\left(\frac{A}{C}\right)$ and *specificity* Sp is equal to $p\left(\frac{\bar{A}}{\bar{C}}\right)$. The choice of a rule quality measure depends on

the specific goal of the mining process. For instance, in a classification task for a 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 optimise sensitivity but not specificity. In other domains, one may want to optimise both sensitivity and specificity simultaneously. For instance, in [2] a genetic approach is considered to optimise the product $Se \times Sp$.

For association rules, *support* and *confidence* have proved their limits since they favour numerous rules which are most often irrelevant. For a rule $A \rightarrow C$ $support = p(A \cap C)$ and $confidence = p\left(\frac{C}{A}\right)$.

Measures of interestingness were motivated by the remark that confidence is not relevant. Indeed, if $p(C/A)$ is greater than a given threshold, but $p\left(\frac{C}{A}\right) < p(C)$, then the rule is not useful. Most interestingness measures generally compare the *a priori* probability $p(B)$ and the *a posteriori* probability $p\left(\frac{C}{A}\right)$. The RI measure from G.

Piatetsky-Shapiro [4] is defined by the expression : $|A| \times (p\left(\frac{C}{A}\right) - p(C))$, the *lift* measure [3] is equal to $\frac{p(C/A)}{p(C)}$. The measure $\frac{p(A \cap C)}{p(A \cap \bar{C})}$ defined by Sebag and

Schoenauer [6] and the *J-Measure* [7] counts negative examples too.

$$J\text{-Measure}(A \rightarrow C) = p(A) \times \left(p\left(\frac{C}{A}\right) \times \log\left(\frac{p\left(\frac{C}{A}\right)}{p(C)}\right) + p\left(\frac{\bar{C}}{A}\right) \times \log\left(\frac{p\left(\frac{\bar{C}}{A}\right)}{p(\bar{C})}\right) \right)$$

This kind of measures tends to quantify the rule relevance by comparison with the situation of independence between the conditional part and the conclusion part. But

rule quality is not reduced to this criterion only. The goodness of a rule cannot be defined in an absolute manner, it depends on specific objectives. For instance quality may involve correlation between condition and conclusion, but generality and coverage too. Or if we are searching for rare events, quality involves rarity and precision.

Thus the questions we address here are : how to measure rules which satisfy multiple criteria simultaneously? are there correlations among confidence, generality, interestingness according to *J-Measure*, or *RI*, *lift*, sensitivity, specificity ? Is it right to use a product of terms like $|A| \times (p(C/A) - p(C))$ or $Se \times Sp$ for optimising the set of solutions?

In this paper, we give some results of a comparison study we have led on these measures. In order to highlight correlations or no-correlations between measures, we have plotted sets of points where each point maps a rule measured according to two measures. We have studied three types of rules : classification rules and association rules extracted by standard algorithms and randomly generated rules. Resulting graphs show the relative behaviour of measures. In the same time, they show where standard algorithms for classification and association are focusing in the rule search space. We have observed apparent antagonisms between measures which suggest that no optimal solution may exist but compromises only. Thus we have implemented an optimisation method to elicit best compromises. Thereafter the presentation focuses on two different cases. Case 1 is related to criteria Se and 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 : the measure of generality $|A|$ and the measure of difference from the independence situation $p(C/A) - p(C)$. These two factors are considered to represent interestingness in the *RI* measure.

3 Correlation among measures

In this initial study, pairs of measures have been selected for comparison. Experiences have been built on data files from the UCI repository [8]. For each pair of measures we have compared rules obtained by running implementations of *C4.5* and *Apriori* and randomly generated rules. We have used the Weka open source system from the University of Waikato [9,10] which proposes an implementation of an advanced version of *C4.5* called *J4.8* and *APriori*. Of course, in order to get consistent results when comparing results to *J4.8*, we have observed rules $A \rightarrow C$ where C is composed of one attribute-value term only. 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 until 4000 rules were generated. Thereafter, we present two results as illustrations of case 1 and case 2 given above in section 2. They have been operated on datasets Dermatology and Zoo respectively. Through case 1, we will notice the relative behaviour of sensibility and specificity. Case 2 illustrates the complementary roles of a measure of generality and a measure of interestingness.

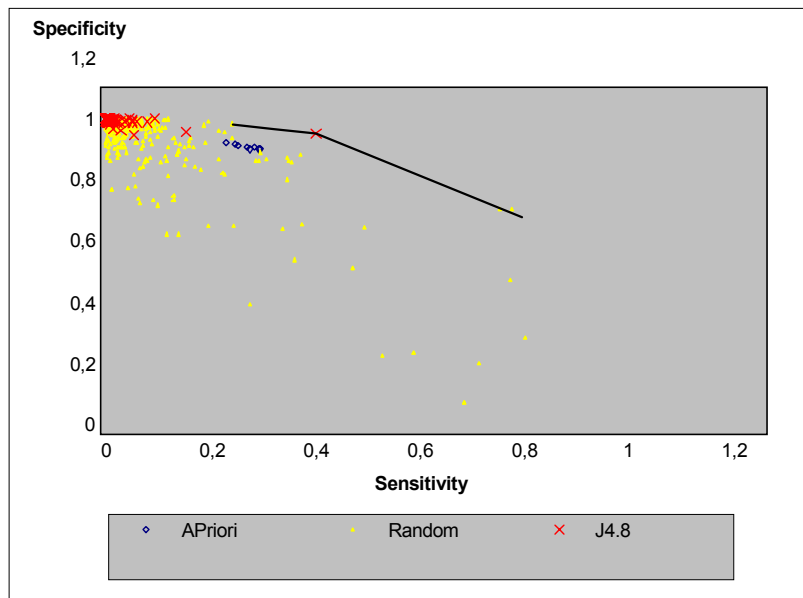


Figure n°1 : Se versus Sp on Dermatology

On Figure n°1, we can observe results from *J4.8*, *Apriori* and random generation for the pair (Se, Sp) on the Dermatology database. We have plotted rules $A \rightarrow C$ where C is a one-attribute-value term. These two measures are more significant than the success rate and the mining process looks for patterns which optimise both of them since Se quantifies the coverage of C by A and Sp quantifies the coverage of \bar{C} by \bar{A} . On Figure n°1 it seems that there is no optimal rule which sensitivity and specificity are both best values : the upper line drawn on the graph crosses over best points and suggests that there is a border. No rules is plotted in the upper right corner. This graph suggests an apparent antagonism between the two measures on this database since there is no point which were best on Se and Sp simultaneously. On this example, when Se is increasing, Sp is decreasing. Thus only compromises may be selected. *J4.8* and *Apriori* finds good rules but there exist other good rules that they ignore. A first idea to obtain best compromises may be to optimise the product $Se \times Sp$ on the set of the rules discovered by *Apriori* or *J4.8*.

There are two drawbacks from this technique. First we can see on Figure n°1, that randomly generated rules give better points in comparison with these standards algorithms. Consequently, techniques that select rules generated according to other criteria's like information gain for *C4.5* or support and confidence for *Apriori*, are not the most efficient. Alternative were explored : for example, M.Fidelis [2] uses an optimisation method that associates extraction and evaluation phases. This method employs an evolutionary techniques in order to optimise $Se \times Sp$.

Then another drawback is that, although the product enables to find rules that standard algorithms miss, we can use more precise techniques. Indeed, there exist good compromises that cannot be found by product optimisation.

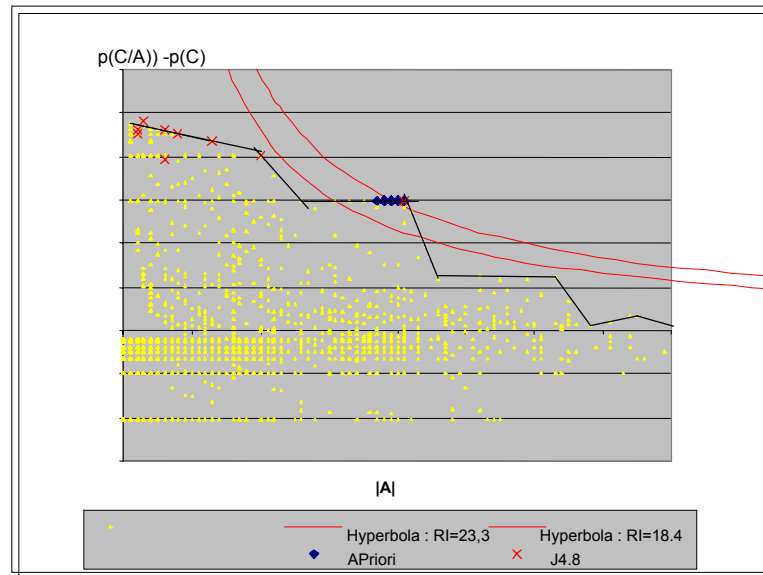


Figure n°2 : $|A|$ versus $conf(r) - p(C)$ on Zoo

In case 2, terms $|A|$ and $p(C/A) - p(C)$ respectively evaluates the size of the condition and the correlation (positive, negative or null) between condition and conclusion. On Figure n°2 drawn from the Zoo database, we check the lack of good solutions in the upper right corner, thus there is an apparent antagonism again. We can observe that the best product is met by a rule determined by *Apriori* (the hyperbola in dotted lines). By defining a range, it is possible to select other rules, mostly generated by *Apriori* and optimising this product in a somewhat different way as shown by the hyperbola in a plain line. However, these hyperbolas help us to visualize that optimising the product does not allow to select other best rules.

Thus, we are faced to a multi-criteria optimisation problem to obtain the whole set of best compromises : we have plotted a set of straight lines which approximate the frontier of best compromises. This observation led us to use a scalar optimisation method to get a better approximation of the set of solutions. Next section presents the scalar method we have implemented.

4 Multi-Criteria Rule Selection

In this section, we propose a solution to approach the set of best solutions among the rule space. This solution is inspired by the graphical interpretation of the figure n°2on

which we have observed that a set of segments may approach the set of best compromises. Thus we have implemented a scalar solution for the multi-measure optimisation by defining different weighted sums of two measures and searching for rules which optimise each sum.

We propose a genetic algorithm (GA) to optimise a function defined as a weighted sum of two quality measures. We have chosen a genetic algorithm for three main reasons :

- GAs evaluation function allow to implement the concept of weighted sum easily,
- GAs yield interesting issues because of their genetic operators (selection, crossover and mutation) that allow to explore efficiently huge search spaces which are frequent in this context
- GAs are potentially able to avoid to get stuck in local optima.

Individual representation

There are two different approaches to extract rules via a GA: the Pittsburgh approach and the Michigan approach [11]. The first one consists in coding several rules inside one individual while according the second one, a rule encodes for one individual only.

In this paper, we adopt the Michigan approach and each individual represents the condition part of a rule *IF condition THEN conclusion*. The conclusion is a one-attribute-value term and this goal attribute is fixed for each run. For discovering several rules predicting different goal attributes, it is necessary to launch several runs, one for each value of the goal attribute. Thereafter we present an individual representation, the fitness function, the selection and genetic operators used and finally the experimental results obtained by application of this algorithm on previous cases 1 and 2.

An individual genome represents a conjunction of attribute-value terms that maps a rule condition. The coding is a position coding that consists in a sequence of genes arranged in the same order that the attributes in the data set. Each attribute-value term ($A_i \text{ op } v$) where A_i is the i_{th} attribute of the data set and v is a value of the A_i attribute domain is coded by a gene. The term *op* is a comparison operator in $\{ =, <=, >= \}$. Each gene i encodes a boolean B_i which indicates whether or not the gene is activated, (i.e. if the attribute is present or not in the rule). So, despite the fact that all genotypes have the same length, rules may have a variable length. The algorithm can handle numeric, boolean and categorical variables. For a categorical one, the operator used is "=". The "<=" and ">=" operators are used for numeric variables.

Fitness function

As mentioned above, our goal is to simultaneously optimise different criteria. This is easily implemented in the GA via the fitness function. GAs are robust algorithms elaborated for optimisation ; this is realized mainly thanks to the fitness function. The fitness value of a rule r is defined as a weighted sum of measures :

$$fitness(r) = \sum_{i=1}^k w_i \times measure_i(r) \text{ such that } \sum_{i=1}^k w_i = 1$$

The fitness function is applied to each individual when the initial population is initialised and each time an individual undergoes a modification due to the crossover or mutation. The evaluation of an individual is realised on the entire data set.

Selection and genetic operators

The algorithm uses the tournament selection with two individuals which consists to randomly choose two individuals in the current population, and then to compare their fitness values. In a picturesque way of seeing, the one which has the best fitness value wins the tournament and will eventually transmit its genetic material at the next generation. This process is repeated P times, with P the size of the population. Such selected individuals are then submitted to genetic crossover and mutation operators.

The algorithm uses a single site crossover (crossover rate is 0.8). A site is randomly chosen and the genetic material located after this site is exchanged between the two parents. The site is chosen in such a way to avoid to generate empty individuals.

The mutation operator modifies rule length. It specifies a rule by the adjunction of a condition if no gene is activated, or it generalizes it by deleting a gene if the rule length is strictly greater than one. The adjunction of a condition consists in the random generation of an operator and a value in the attribute domain associated to the gene. When deleting a gene mutation generalizes a rule and specializes it by the adjunction of a condition generated in a random way. The mutation operator is applied with a rate of $1/m$, where m the length of the genotype.

Experimental results

Our scalar method has been tested on the datasets Dermatology and Zoo databases. Dermatology contains *366* descriptions of skin diseases according to *33* categorical attributes and *one* numeric attribute. This dataset was used for classification on the goal attribute *disease*. Zoo contains *101* descriptions of *7* animal types according to *17* boolean attributes and *one* numeric attribute. This dataset was used for classification on the goal attribute *type*.

In each of these tests goal attributes are discrete and individuals represent rules $r : A \rightarrow C$ where C is an attribute-value term defined on the goal attribute. The population size is fixed to *100* individuals. A run of the GA consists in *100* generations. We realized 10 runs for the same GA to optimise the fitness function defined by a couple of weights.

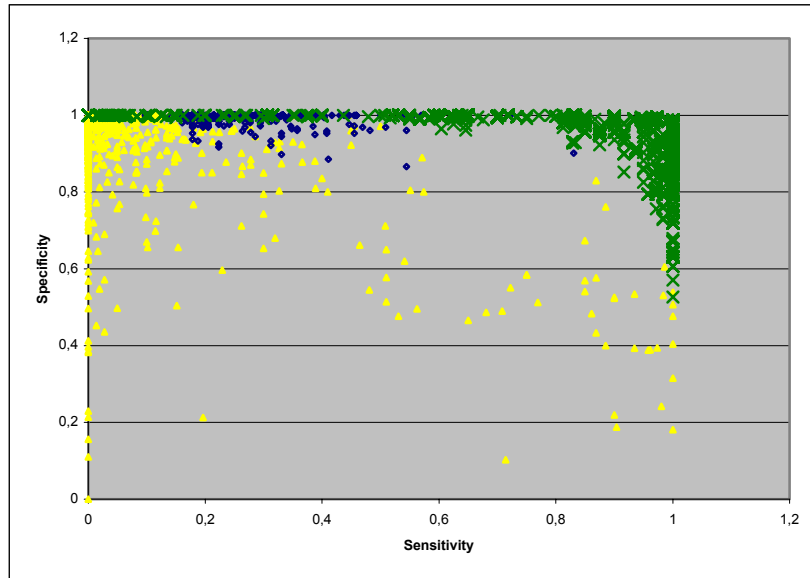


Figure n°3 : *GA_Optimise*(Se,Sp) results on Dermatology

The results obtained via the GA are produced by the following algorithm *GA_Optimise*.

```

algorithm GA_Optimise(x,y)
// for a pair c=(x,y) of measures
begin
  c(x',y') <-- normalize(c(x,y))
  for each value vk of the goal attribute do
    for each w ∈ {0.1,0.2,...,0.9} do{
      launch 10 times the GA with
      fitness= w*c.x' + (1-w)*c.y'
      and
      vk to predict}
      // the best individual is kept
      // in the final population
    end
  end
end

```

We have chosen empirically nine different couples of weights to reduce the execution duration. Figures n°3 and n°4 show the results obtained by *GA_Optimise* respectively on pairs (Se,Sp) and (|A|, p(C/A)-p(C)) on which we have detected possible antagonism according Figures n°1 and 2.

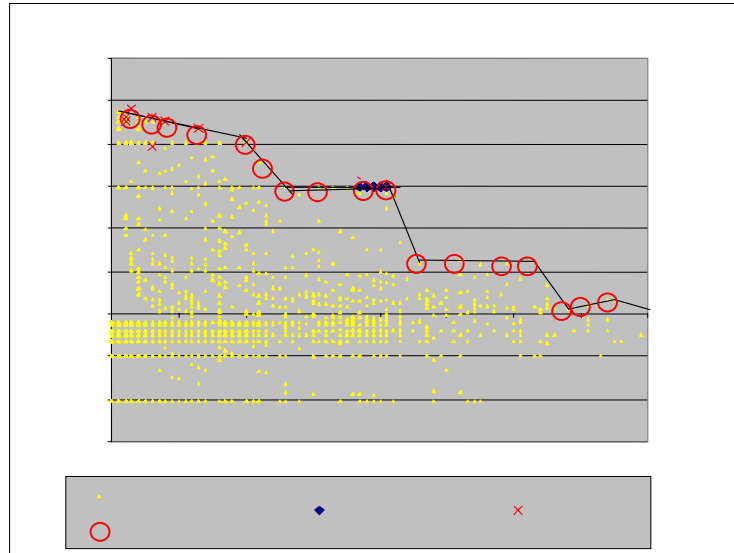


Figure n°4 : $GA_Optimise(|A|,p(C/A)-p(C))$ results on Zoo

Rules extracted by *J4.8* and *APriori* and generated randomly are figured together with GA rules. Figure n°3 shows the emergence of good rules extracted by *GA_Optimise* and ignored by standard algorithms. The comparison with Figure n°1 shows the evolutionary scalar method efficiency in the search space. Indeed rules in the upper right corner have emerged with the scalar method while they were not found before. Thus *GA_optimise* is able to find best rules since it simultaneously evaluates both criteria of sensibility and specificity. The real advantage of *GA_optimise* is illustrated by Figure n°4 on case 2.

On Figure n°4 GA rules are figured by circles. One can observe that circles are located on the upper limit of the graph but are not located in the upper right corner. The scalar method does not find absolute good rules and confirms by that the antagonism previously appearing between the two measures. The comparison between Figures n°2 and n°4, illustrates the advantage of the scalar method which allows to obtain a best approximation of the set of compromises on both measures $|A|$ and $p(C/A)-p(C)$. Indeed, each run of the GA optimising a different weighted sum of the two measures identifies solutions on the neighbourhood of the set of best solutions.

5 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 even domain-dependent qualities too. Thus, we have addressed the problem of selecting rules according multiple criteria. A comparative study on different measures and standard mining algorithms has highlighted the limits of these algorithms and has shown phenomena of antagonism between measures. These observations have led us to look for a multi-criteria optimisation method allowing to find a set of best compromises rather than an optimal one. An evolutionary algorithm has been chosen since this kind of stochastic methods is able to explore large search spaces and allows to define easily different evaluation function. A scalar method has been tested and in particular it has extracted rules which were ignored by standard algorithms.

Since scalar optimisation is not the most efficient to afford an exhaustive search in the set of best compromises, our current activities on this subject are oriented towards more robust optimisation methods.

References

- [1] R. Agrawal and A. Srikant. Fast algorithms for mining association rules. Proc. VLDB'94, pp 487-499.
- [2] M.V. Fidelis, H.S. Lopes and A. A. Freitas. Discovering comprehensible classification rules with a genetic algorithm. Proc. Congress on Evolutionary Computation (CEC-2000), pp 805-810. La Jolla, CA, USA. July 2000.
- [3] International Business Machines, IBM intelligent Miner, User's guide, version 1, release 1, 1996.
- [4] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. G. Piatetsky – Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases. MIT Press. 1991.
- [5] J. R. Quinlan. C4.5 : Programs for Machine Learning. Morgan Kaufmann, 1993.
- [6] M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases.. Proc. of the European Knowledge Acquisition Workshop, EKAW'88. 1988.
- [7] P. Smyth and H. M. Goodman. Rule induction using information theory. Knowledge Discovery in Databases. G. Piatetsky – Shapiro and W. J. Frawley, editors, MIT Press. 1991.
- [8] UCI Machine Learning, <ftp.ics.uci.edu/pub/machine-learning-databases>.
- [9] Weka Software, University of Waikato, New Zealand, <http://www.cs.waikato.ac.nz/ml/weka/>
- [10] I. H. Witten and E. Frank. Data Mining. Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers. 1999.
- [11] D. E. Goldberg. Genetic algorithms in search, optimization and machine learning. *Addison Wesley*, 1989.