Interactive Search of Rules in Medical Data Using Multiobjective Evolutionary Algorithms

Daniela Zaharie
West University of Timisoara
4 V. Parvan, Blvd.
Timisoara, Romania
+(40)256 592157
dzaharie@info.uvt.ro

Diana Lungeanu
University of Medicine and Pharmacy
2 E. Murgu, Sq.
Timisoara, Romania
+(40)256 490288
dlungeanu@umft.ro

Flavia Zamfirache
West University of Timisoara
4 V. Parvan, Blvd.
Timisoara, Romania
+(40)256 592157
zflavia@info.uvt.ro

ABSTRACT

In this work, we propose an approach for evolving rules from medical data based on an interactive multi-criteria evolutionary search: besides selecting the set of criteria and the sets of potential antecedent and consequent attributes, the user can also intervene in the searching process by marking the uninteresting rules. The marked rules are further used in estimating a supplementary optimization criterion which expresses the user's opinion on the rule quality and is taken into account in the evolutionary process.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms, Human Factors

Keywords

Rules mining, multiobjective optimization, evolutionary algorithms, interestingness measures, interactive search

1. INTRODUCTION

Discovering new and useful knowledge from medical data represents a challenge for any data mining task, due to the heterogeneous nature of medical data (usually consisting of mixed attributes, i.e. nominal, numerical, and logical, with many erroneous or missing values) and to the requirements to express the knowledge in a medically comprehensible form. The final aim is assisting the medical specialists in making decisions, so the data mining tools should help them explore large volumes of medical facts stored in databases and extract meaningful rules upon which they can rely in both medical research and clinical hypothesis formulation [14].

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.

An easily understandable manner of expressing hypotheses extrapolated from data is represented by rules in the form:

IF "some conditions on the values of predicting attributes are true" THEN "some conditions on the goal attributes are true"

If there is only one goal attribute and it specifies a class, then we discuss about a classification rule, expressing the possibility that data satisfying the antecedent (IF) condition belong to the class specified in the consequent (THEN) part. When the goal attributes do not express a class, then we deal with prediction rules, expressing hypotheses on the dependence between the antecedent and consequent parts of the rules. Finally, when the potential sets of antecedent and consequent attributes are not previously established, we investigate general association rules expressing co-occurrence of different attribute values. Discovering and selecting rules in data is a search process usually guided by several measures based on which their potential quality and usefulness is evaluated: (i) accuracy measures (quantifying the ability of the rule to describe the data); (ii) comprehensibility measures (expressing the readability and understandability of rules); and (iii) interestingness measures (quantifying the potential to provide new, previously unknown knowledge). These measures are usually conflicting, i.e. an accurate rule is not necessarily interesting or easy to read, thus the searching process has to be multicriterial and dozens of such measures have been proposed and investigated [3, 13].

Evolutionary algorithms (EAs) proved to be valuable instruments in data mining [8] and a significant number of works describe the use of EAs in discovering rules from data [2, 7, 11, 12] or in post-processing the set of rules previously extracted by non-evolutionary approaches [9, 10]. Except for the work of Ishibuchi and Yamamoto, they treat the multi-criterial character of the search by aggregating all criteria in a single one through a pre-specified aggregation function (e.g. a product or weighted sum). Approaches based on multi-objective evolutionary algorithms (MOEAs) have also been proposed [1, 6].

However, these approaches fail to take into account the user: no set of quality criteria can be exhaustive, so the user should be involved in the search process. In this work, we propose an approach for evolving rules from data based on an interactive search: besides selecting the set of criteria and the sets of potential antecedent and consequent attributes, the user can also intervene

in the searching process by marking the uninteresting rules. They are further used in estimating a supplementary criterion which expresses the user's opinion on the rule quality and is taken into account in the evolutionary process itself as all the other criteria.

2. EVOLUTIONARY APPROACHES IN RULES MINING

2.1 Designing an EA for Rules Mining

When designing an EA for rules mining, one has to take into account at least the following aspects: (i) representation of rules; (ii) initialization of the population; (iii) recombination and mutation operators; (iv) rules evaluation and selection.

There are two main approaches for rules representation: "Pittsburgh" and "Michigan". In the former approach, each element of the population is a set of rules, thus it deals well with the interaction between rules and is appropriate when looking for sets of rules defining a classifier. On the other hand, the evolutionary operators are complex and rather difficult to be implemented [8]. In the latter approach, each element of the population encodes one rule, therefore all elements have the same structure and the evolutionary operators can be more easily implemented. Although it does not deal well with the rules' interaction, the "Michigan" approach is the most frequently used, especially for prediction and association rules, where the interaction is not as critical as it is in the case of the classification ones.

The evolutionary process starts from the initial population, so choosing appropriate initial rules compatible with the actual data is important. Therefore, most EAs for rules mining start with randomly generated rules, while still trying to involve values present in the data to be mined. The mutation and recombination operators are adapted in order to ensure the rules consistency.

The success of an evolutionary rules mining process is highly dependent on the quality measures used to evaluate the population elements and on the strategy of selecting the elements to be transferred in the next generation. Since the quality of a rule depends on several criteria there are a large number of variants to compute the fitness value.

2.2 Previous Work

During the last decade a significant number of evolutionary algorithms were designed to assist the process of extracting rules from data. The proliferation of the evolutionary approaches in rules mining is motivated by the fact that EAs can deal well with continuous numeric data and can be easy adapted to solve tasks arising at different stages of a rules mining process.

For instance, Mata et al [11] proposed an evolutionary algorithm for finding the frequent item sets in numeric databases. The advantage of this approach over classical non-evolutionary techniques is that it does not require a previous discretization of continuous data. The characteristic of this algorithm is that the frequent item sets are iteratively discovered by guiding the searching process through a penalization mechanism of data instances already included in a subset. The fitness function is obtained by aggregating several terms related with: the rule's support; the amplitude of the intervals of values corresponding to attributes; the number of attributes in the item set; and the penalization expressing the ratio of data covered by the current rule which are also included in other item sets.

Other variants, as those proposed by Gopalan et al [9], use an EA to post-process a rule set extracted using different techniques. The post-processing aims at discovering the accurate and interesting rules in large sets of classification rules. The evolved structures correspond to sets of rules, as in the "Pittsburgh" approach, and the selection process consists of two stages: (i) accurate sets of rules are selected; then (ii) from these accurate sets of rules, the most interesting ones are selected. In [10], the EAs are also used for post-processing an existing set of rules, but they deal with fuzzy association rules and use a multi-objective approach.

Evolutionary algorithms are also used to extract classification rules [7] and prediction rules [12] (satisfying the comprehensibility and interestingness requirements) directly from data. The rules' quality is computed by aggregating several accuracy, comprehensibility and interestingness measures, thus the EA has to deal with a single-objective optimization problem.

Recently, Pareto-based multi-objective algorithms have been used to extract fuzzy association, numeric association, or classification rules [1, 6]. However, their common problem is that a single run of the algorithm leads to a set of rules which can be quite large, especially when the number of criteria is large.

In order to apply such a technique to medical data, one should pay special attention to the nature of data (usually containing mixed attributes and a significant ratio of missing values) and to the choice of accuracy and interestingness measures.

3. MEASURES FOR EVALUATING THE RULES QUALITY

Let us consider rules having the following structure:

$$R : (AT_1,...AT_k) \rightarrow (CT_1,...,CT_l) \quad (1)$$

where *AT* denotes an antecedent term and *CT* denotes a consequent term. Each term involves one data attribute and is a triplet $\langle a, op, value \rangle$ where *a* is an attribute, *op* is an operator (equal, different, in, not in, less than, greater than) and *value* is a possible value or a set of values for the attribute. Each term evaluates to a Boolean value, thus the rule described in (1) can be read as:

IF $AT_1,...,AT_k$ *are all true THEN* $CT_1,...,CT_l$ *are all true* (2)

A data instance $(a_1,...,a_n)$ satisfies (or is covered by) a given rule if all terms (antecedent and consequent) are true for the attribute values in that instance. If the antecedent terms are all true but there exists at least one consequent term which is false, then the data satisfy only the antecedent part of the rule. Similarly, if all consequent terms are true but at least one antecedent term is false then the data satisfy only the consequent part of the rule. Let us denote by *A* the event that the antecedent part of the rule is satisfied (disregarding the satisfaction of the consequent) and by *C* the event that the consequent part of the rule is satisfied (disregarding the satisfaction of the antecedent). Then *P(A,C)* will be the probability that both the antecedent and consequent parts are satisfied, *P(A)* will be the probability that the antecedent part is satisfied, and *P(C)* the probability that the consequent part is

satisfied. The negation $\overline{A}$ will denote the event corresponding to the case when the antecedent part is not satisfied. We denoted in a similar manner the negation of C and the corresponding probabilities.

3.1.1 Accuracy measures

Accuracy measures reflect the likelihood of rules (given the actual data) and the most frequently used are: rule support (Supp), confidence (Conf), accuracy (Acc), specificity (Spec) and sensitivity (Sens). They are defined as follows:

$$\begin{aligned} &Supp = P(A,C) \\ &Conf = P(A,C)/P(A) \\ &Acc = P(A,C) + P(\overline{A},\overline{C}) \\ &Spec = P(\overline{A},\overline{C})/P(\overline{C}) \\ &Sens = P(A,C)/P(C) \end{aligned} \quad (3)$$

The accuracy, specificity and sensitivity are mainly used in the case of classification tasks, therefore they can also be defined by using the elements of the confusion matrix (true positive/negative cases, false positive/negative cases). For association and prediction rules, the typically used measures are the support and confidence. In the case of medical rule mining, these measures have to be treated with caution as high support or even high confidence rules are not necessarily interesting from a medical point of view.

3.1.2 Comprehensibility measures

The readability of a rule is usually related to its length, i.e. the number of related terms, therefore we used as a comprehensibility measure:

$$ch = 1 - \frac{l+k}{n} \quad (4)$$

where n is the maximal number of terms (in the antecedent and the consequent part) and $l+k$ is the effective number of terms in the rule.

3.1.3 Interestingness measures

There are more than 40 objective measures for evaluating the interestingness of a rule [13], so choosing appropriate measures for the actual data characteristics is not a trivial task, as Carvalho et al clearly illustrated in their paper, too [3]. They analyzed the correlation between objective interestingness measures and the real human interest evaluated by experts from each domain, and proposed a ranking of objective quality measures.

Not quite surprisingly, different rankings were obtained for different datasets. For the medical data set they tested, the top three measures were: Phi-coefficient (Φ), odds ratio (OR) and cosine measure (cos), as described in equations (5).

$$\begin{aligned} &\Phi = \frac{P(A,C) - P(A)P(C)}{\sqrt{P(A)P(C)(1-P(A))(1-P(C))}} \\ &OR = \frac{P(A,C)P(\overline{A},\overline{C})}{P(A,\overline{C})P(\overline{A},C)} \\ &\cos = \frac{P(A,C)}{\sqrt{P(A)P(C)}} \end{aligned} \quad (5)$$

The cosine measure is similar with the interest (lift) measure, as described in equation (6).

$$lift = \frac{P(A,C)}{P(A)P(C)} \quad (6)$$

Ohsaki et al [13] presented a similar study, exclusively oriented towards medical data. A large set of measures (41) were evaluated according to some metacriteria expressing the relationship between the objective measure value given to a rule and the quality label assigned to the same rule by a medical expert. The top three measures obtained by combining the rankings corresponding to two medical sets (one on meningitis and the other on hepatitis) were: accuracy, peculiarity, and uncovered negative (UN). As peculiarity can be estimated only in the case of discrete attributes, we included in our analysis only the last one as a measure of interestingness:

$$UN = P(\overline{A},\overline{C}) \quad (7)$$

We can see that UN is the difference between accuracy and support, so maximizing this criterion leads to the maximization of the accuracy and minimization of the support. Thus it favours rules that are not necessarily of high support, but could be interesting.

4. AN INTERACTIVE MOEA FOR RULES MINING

We propose to involve the user in the search process, as a predefined aggregation of quality criteria is difficult to find and, moreover, it has been suggested that users can also change their opinion on the rules' quality during the evaluation process itself [13]. The approach we propose is based on a multi-objective evolutionary algorithm having the general structure described in Figure 1.

4.1 Characteristics of the Evolutionary Algorithm

4.1.1 Rules Encoding

Each element (chromosome) of the population corresponds to a rule and it consists of a list of components (genes) corresponding to all attributes in the data set. Each component consists of three fields: (*presence flag, operator, value*).

The *presence flag* is a binary value specifying whether the corresponding attribute is involved in the rule (either in its antecedent or the consequent part). In case of binary attributes, this is the only important field.

The *operator* allows specifying the condition the attribute should satisfy. We used two possible operators for each type of attributes. In the case of numerical attributes the possible

operators are ≤ (coded by 0) and > (coded by 1). For the categorical attributes the operators are = (coded by 1) and ≠ (coded by 0).

Generic MOEA for rules extraction	
1:	Initialize a population of *m* rules
2:	Evaluate the population
3:	REPEAT
4:	Evaluate the rules in the current population
5:	Generate *m* new rules by crossover
6:	Apply mutation to rules obtained by crossover
7:	Evaluate the new elements
8:	Join the old and the new populations
9:	Select the "best" *m* rules from the joined population
10:	UNTIL "a stopping condition is satisfied"

Figure 1. General structure of the evolutionary algorithm.

The *value* field contains the value associated to the attribute. In the case of numerical attributes, the terms can be of interval type (e.g. *a in [min,max]*) and the *value* contains the lower and the upper limits of that interval. The operator field is also differently interpreted in that situation (e.g. 0 encodes the operator *not in* and 1 encodes the operator *in*).

In all cases, an element is a fixed-length list with mixed (binary, integer, and real) values. In the following, the number of attributes is denoted with *n*.

The difference between the antecedent and consequent attributes is made only in the evaluation of an element. In case of the classification rules, the class attribute is not included into the population elements, all attributes being predictive.

4.1.2 Reproduction Operators

During each generation, a new population is constructed by crossover and mutation from the current one. By crossover, a new rule is constructed starting from two randomly selected rules from the current population. In case of rules containing only terms of inequality type, the crossover procedure can be described as following:

(i) if the attribute is absent from the both parent rules, it will be absent from the generated rule, as well;

(ii) if the attribute is present in only one parent rule, its operator and value field are transferred to the new rule;

(iii) if the attribute is present in both rules and satisfies the same type of condition (the operator field has the same value in both rules) then it is transferred to the new rule: for numerical attributes, the new value is the average of the values corresponding to the parent rules; for nominal attributes, one of the parents' values is just randomly taken;

(iv) if the attribute is present in both rules and it satisfies different conditions, then the triplet to be transferred into the new rule is taken from the parent rule which is better with respect to one of the evaluation criteria (in our experiments we used the first criterion specified by the user).

The mutation has the role of modifying the rules obtained by crossover. For each attribute, mutation is applied with a given probability (e.g. $p_m=0.1$ *or* $p_m=1/m$) and it can affect one of the fields (i.e. presence flag, operator or value) and only one at each mutation step. By switching the presence flag, some attributes can be inserted or removed from the rule, thus leading to either a more general or a more specific rule. By changing the operator field, one changes the condition the attribute should satisfy. The mutation of the value field consists in choosing a new value based on a uniform selection from the range of values corresponding to the attribute. If the new element generated by crossover and the mutation is not valid (e.g. it does not contain any antecedent or consequent term), then a repairing rule is applied (e.g. a randomly generated antecedent or consequent term is introduced).

In case of rules containing terms of interval type, we used the recombination operator described in [1].

4.1.3 Selection and Archiving

After a new population is created by crossover and mutation, a selection step (typical to MOEAs) is applied. Our selection strategy is similar to that used in NSGA-II [5], meaning that the elements in the joined population (parents and offsprings) is ranked based on the non-domination relationship. A rule is considered as non-dominated, with respect to rules in a given set, if no other rule in that set is better with respect to all criteria. All the elements that are non-dominated with respect to all elements in the joined population belong to the first nondomination front and have the rank 1. Subsequently, the nondominated elements in the population obtained by ignoring the elements of rank 1 belong to the second nondomination front and so on. The *m* elements corresponding to the new generation are selected from the *2m* ranked elements in their ranks' increasing order. For stimulating the diversity of the resulting Pareto front, a crowding distance is used as a second selection criterion: from two elements having the same rank, the one with a larger crowding distance (suggesting that it belongs to a less crowded region) is selected. The crowding distance can be defined in either the objective or the decision variables space.

A particular characteristic of our approach is related to the crowding distance between rules. We analyzed two types of distances, one expressing the structural differences between rules and another expressing the difference between the data subsets covered by the rules. In the case of two rules $R=(t_1,...,t_n)$ and $R'=(t'_1,...,t'_n)$, the structural distance is defined as follows:

$$d_S(R,R') = \frac{1}{n}\sum_{j=1}^{n} d_j(R,R'),$$

$$d_j(R,R') = \begin{cases} 0 & \text{if } p_j = p'_j, o_j = o'_j \\ 1 & \text{if } p_j = p'_j, o_j \neq o'_j \\ 2 & \text{if } p_j \neq p'_j \end{cases} \quad (8)$$

where p_j denotes the presence flag and o_j denotes the operator corresponding to attribute *j*. Thus two rules are considered to be identical from a structural point of view if they contain the same

attributes and the terms have identical associated operators. The distance related with the rule coverage is defined as the cardinal of the subset of data which are either covered by the first rule but are not covered by the second rule, or are covered by the second rule and are not covered by the first one. Thus the cover-based distance is:

$$d_C(R, R') = card(\text{cover}(R)\Delta\text{cover}(R')) \quad (9)$$

where cover(R) is the set of data instances which satisfy the rule (are covered by the rule) and Δ denotes the symmetrical difference between two sets.

After a given number of generations, an archive of nondominated elements is constructed. Not all non-dominated elements from the current population are transferred in the archive, but they are filtered such that both the structural and the cover-based distances between any two elements of the archive are larger than a given threshold (in our analysis we used 0.01).

4.2 User Guided Evolutionary Search

An interactive search allows the user to interfere with the evolutionary process in order to guide it towards interesting regions of the search space. The overall idea of the interactive process of rules evolving is illustrated in Figure 2. In the interactive variant, the search process consists of several steps; at each one, the population is evolved for a given number of generations and the archive of the selected non-dominated rules is provided to the user together with all the objective measures computed for the testing dataset (measures not necessarily limited to those used as criteria in the optimization process). In our implementation, we used the following list of measures: support, confidence, accuracy, specificity, sensitivity, comprehensibility, Phi-coefficient, odds ratio, lift, Piatesky-Shapiro, Jaccard, kappa-coefficient [3] and also uncovered negative and relative risk [13].

Based on these criteria and on a subjective evaluation, the user can decide whether there are uninteresting or incomprehensible rules. Then (s)he can mark these rules and proceed to the next step of the search. The effect of marking the undesirable rules is twofold: firstly, the population elements corresponding to the marked rules are replaced with randomly initialized elements; secondly, the marked rules are added to a list (L_p) of prohibited rules. This list is used to compute a supplementary optimization criterion which expresses the user's evaluation. For a rule *R,* this criterion is computed as the distance between *R* and the list of prohibited rules:

$$ue(R) = \min\{d(R, R'); R' \in L_P\} \quad (10)$$

The distance between rules can be either the structural distance (eq. (8)) or the cover-based distance (eq. (9)) introduced in the previous section. By using the user's evaluation as a quality criterion, rules "similar" to those marked as uninteresting have little chance to evolve and survive during the next steps of the evolution. On the other hand, employing the user's evaluation in the search can redirect it towards different regions of the searching space, thus leading to the discovery of new rules. This diversity enhancing effect is analyzed in Section 5.2. A possible drawback of using a supplementary criterion is that it usually leads to a larger number of nondominated elements.

Another interactive variant of MOEAs is presented in [5], but it is not based on introducing new optimization criteria; it uses other techniques (non-evolutionary) to improve the current Pareto front or to focus the search towards a given region of the front.

Interactive variant

1: Initialize a population of *m* rules
2: Evaluate the population
3: FOR step:=1,maxStep DO
4: Execute the MOEA (lines 3-10 in Figure 1)
5: Construct the archive of rules
6. Evaluate the rules in the archive for a test dataset
7: Visualize the rules' archive
8: Get the user evaluation on the rules in the archive
9: Process the rules marked by the user as uninteresting:
- Replace the corresponding elements from the population with newly initialized elements
- Add the marked rules to the list of prohibited ones

10: Re-evaluate the current population by taking into account the user evaluation criterion (eq. (10))
11: ENDFOR

Figure 2. General structure of the interactive variant of the rules' extraction algorithm.

5. EXPERIMENTS IN MEDICAL RULES MINING

The interactive variant described in the previous section was implemented such that the user can select the following elements:

- *Rules type.* Both classification and prediction rules can be evolved. In case of classification ones, each run of the algorithm leads to rules corresponding to one class and only the antecedent terms of a rule are encoded, the consequent being implicit.
- *Lists of attributes* to be included in the antecedent/consequent parts of the rules.
- *Criteria* to be used in the searching process. The user can choose an arbitrary subset of measures from those available (mentioned in the previous section).
- *Parameters of the evolutionary algorithm*. The main parameters the user can set are: population size; crossover and mutation probabilities; number of generations between two consecutive user evaluation steps.

The approach was tested for the medical datasets from the UCI repository (http://mlearn.ics.uci.edu/MLRepository.html) and for a set of obstetrical data collected during 2006 in a hospital of obstetrics-gynaecology. The aim of the experiments was twofold: to validate the ability of the evolutionary approach to discover accurate rules, and to analyze the impact of the user's intervention in the searching process.

5.1 Validation in the Case of Classification Rules

In order to validate the ability of the implemented multi-objective evolutionary algorithm to extract reliable rules, we firstly tested it in the case of classification problems. The approach we followed was based on the idea of evolving rules corresponding to one class. Therefore, only the terms in the antecedent part of the rules were evolved. As a class cannot be described by a single rule (usually requiring a set of rules), the MOEA provides a set of reciprocally non-dominated rules, which can be interpreted as describing the class itself (even if it does not cover the entire class). So the rules were evaluated both individually and as a set, using a five-fold cross-validation approach. In all tests, the population size was set to 50, the maximal number of generations to 100, and the mutation probability to 0.1.

The results in Table 1 were obtained for Pima Indians Diabetes data set, based on two optimization criteria: accuracy (Acc) and uncovered negative (UN). These results are comparable with those obtained by applying other rule-based classifiers (Table 2) implemented in the Weka data mining tool (http://www.cs.waikato.ac.nz/ml/weka/): simple rules classifiers (ZeroR, OneR), conjunctive rules classifier (CR), decision table majority classifier (DT), propositional rule learner based on repeated incremental pruning (JRIP), nearest neighbor-like classifier with non-nested generalized exemplars (NNge), partial decision trees (PART).

Rules having high accuracy do not necessarily have a high uncovered negative (UN) value; for instance, the highest accuracy rule obtained when using fold 5 of the data set was:

> "*IF (Plasma glucose concentration>127, 2-Hour serum insulin (mu U/ml)<599, Body mass index >29) THEN class=diabetes*"

A rule with a high value of UN was:

> "*IF (Plasma glucose concentration>155, 2-Hour serum insulin (mu U/ml)<613, Body mass index >29, Triceps skin fold thickness<43, Diabetes pedigree function >0, Age <58) THEN class=diabetes*".

Table 1. Measures associated to the set of rules evolved from the Pima data set . Optimization criteria: Accuracy (Acc), Uncovered Negative (UN). Rules containing terms of the inequality type.

Fold	No. rules	Testing set					Training set	
		Acc	Spec	Sens	UN	Lift	Acc	UN
1	10	0.73	0.86	0.50	0.55	1.87	0.78	0.65
2	10	0.73	0.83	0.55	0.53	1.82	0.75	0.65
3	10	0.71	0.88	0.40	0.57	1.84	0.76	0.65
4	6	0.74	0.95	0.35	0.62	2.28	0.75	0.64
5	12	0.77	0.85	0.62	0.55	1.98	0.77	0.65
Avg	9.6	0.73	0.87	0.48	0.56	1.95	0.76	0.64
stdev	2.1	0.02	0.04	0.10	0.03	0.19	0.01	0.004

Table 2. Results obtained by other rule-based classifiers for the Pima dataset

	Zero R	CR	DT	JRIP	NNge	OneR	PART
Acc	0.65	0.71	0.72	0.75	0.74	0.72	0.74

We also analysed the influence of the rules' type by applying another variant of the algorithm to the same set of data: in the case of numerical attributes, we used terms of interval type and the recombination operator proposed in [1]. The corresponding results are presented in Table 3: no significant differences were identified between these two variants. Examples of rules, containing terms of interval type, evolved using the variant of the algorithm based on the operators described in [1] are:

> "*IF Plasma glucose concentration in [156,196], Body mass index in [8.33,58.64] THEN class=diabetes*" (high accuracy rule)

> "*IF Plasma glucose concentration in [156,188], Number of pregnancies in [3,14], 2-Hour serum insulin (mu U/ml) in [87,730], Body mass index in [7.9,59.74] THEN class=diabetes*" (high uncovered negative measure)

Table 3. Measures associated to the set of rules evolved from the Pima data set . Optimization criteria: Accuracy (Acc), Uncovered Negative (UN). Rules containing terms of the interval type.

Fold	No. rules	Testing set					Training set	
		Acc	Spec	Sens	UN	Lift	Acc	UN
1	7	0.70	0.83	0.46	0.53	1.69	0.74	0.65
2	9	0.69	0.83	0.44	0.53	1.66	0.74	0.64
3	8	0.70	0.94	0.27	0.61	2.03	0.70	0.65
4	6	0.76	0.93	0.45	0.60	2.23	0.74	0.65
5	12	0.73	0.94	0.35	0.61	2.19	0.73	0.65
Avg	8.4	0.71	0.89	0.39	0.57	1.96	0.73	0.64
stdev	2.3	0.02	0.05	0.08	0.04	0.27	0.01	0.004

5.2 Impact of User Interaction

In order to gather preliminary information on the impact of the user's intervention in the searching process, we conducted an experiment based on three scenarios:

(i) without user interaction;

(ii) with user interaction, while using the structural distance to evaluate the user criterion, *ue*;

(iii) with user interaction, while using the cover-based distance to evaluate the user criterion, *ue*.

In each case, the evolutionary process was initially guided by only one optimization criterion (accuracy), so in the case of the first scenario only this criterion influenced the search. For the other two scenarios, the user's evaluation criterion influenced the

search process, starting with the second step. In all cases, 10 steps were run starting from the same initial population and, during every step, the population was evolved for 100 generations. At the same time, at each step, the user's intervention (for scenarios (ii) and (iii)) consisted in marking all rules as uninteresting. This was an extreme user intervention, whose effect was guiding the searching process towards different regions in the rules' space and led to variations in the values of the quality measures associated to the rules.

Figures 3 and 4 illustrate the evolution of two interestingness measures (uncovered negative and lift) for the three scenarios described above. As expected, marking some rules as uninteresting and including them in the list of prohibited rules allowed discovery of new rules with different values for the measure of interestingness.

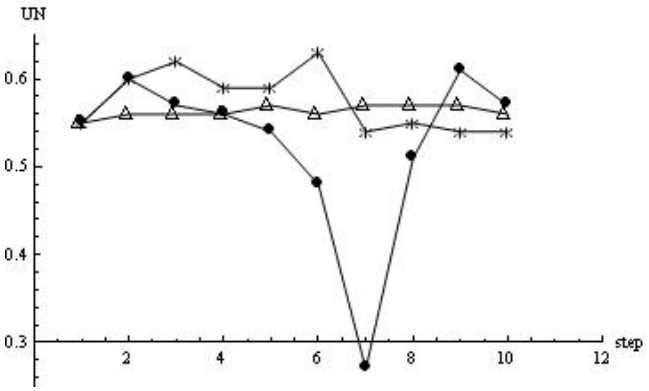

Figure 3. Evolution of the Uncovered Negative (UN) measure: (i) without user interaction (triangles); (ii) with user interaction and structural distance (points); (iii) with user interaction and cover-based distance (stars).

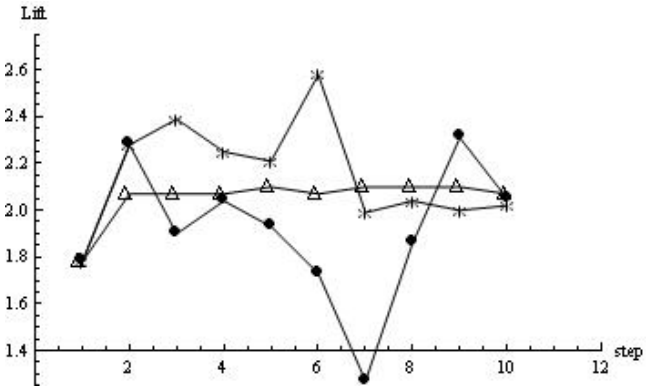

Figure 4. Evolution of the Lift measure: (i) without user interaction (triangles); (ii) with user interaction and structural distance (points); (iii) with user interaction and cover-based distance (stars).

As a further step, we tried to analyze the possibility to explore the rules' space in the case of a set of real obstetrical data containing information about mothers: constitutional characteristics, health status, and the gestational age at the birth moment. The final aim of our investigations was to identify the risk factors for preterm birth and to explore various hypotheses concerning the relationship between the characteristics of the mother and the birth outcome. The set of data contained 2686 records for the births occurred during the year 2006 at one regional hospital of obstetrics-gynaecology. The records corresponded to two main classes: the pre-term birth (370 records, representing 13.77%) and the on-term birth (2316 records, representing 86.23%). Each record contained 63 attributes describing different characteristics of mothers and their new-born babies. The percentage of missing values for the attributes was 23%, which generated difficulties in applying the classification and prediction methods. We considered that records with missing values for the attributes involved in a rule did not match that rule, thus they were ignored when computing the probabilities involved in the rule evaluation measures. In order to limit the size of the search space and of the set with non-dominated rules, we limited our search to the prediction rules having attributes related to the constitutional characteristics of the mother in their antecedent part (e.g. height (h), body mass index before pregnancy (BMI), abdominal circumference (AC) and weight gain (WG) at the birth moment), while the consequent part was the gestational age (GA). Our initial expectation was to find rules having in the consequent part terms of the form: "*Gestational age < a value near or less than 37 weeks*" (in order to catch the pre-term birth cases). Therefore, we started the searching process having in mind the idea of considering as uninteresting the rules with the operator "greater than" in their consequent parts. As optimization criterion, we employed the product between specificity and sensitivity. During the first step, the optimization problem was single-objective, but during the next steps the user's evaluation was taken into account and the problem became a bi-objective one.

Table 4. Evolution of the number of the discovered rules in the case of obstetrical dataset

Step	No. rules	No. rules TC="GA>val"	No. rules TC="GA<val"	User action
1	1	0	1	Mark all rules
(R1) *IF (h>165.3, BMI>7.8, WG<4.36) THEN (GA<22.3)* Supp=0.001,Conf=1,Spec=0.99,Sens=0.99,Lift=538,UN=0.98				
2	1	1	0	Mark all rules
3	3	3	0	Mark all rules
4	5	5	0	Mark all rules
5	2	0	2	-
(R2) *IF(BMI>10.5, AC>95.3,WG<6.4) THEN (GA<25.35)* (Supp=0.001,Conf=0.03, Spec=0.93,Sens=0.5, Lift=9.9,UN=0.93				
(R3) *IF(BMI>1.42,WG<2.32) THEN (GA<21.3)* (Supp=0.001,Conf=0.33,Spec=0.97,Sens=0.97,Lift=179,UN=0.9)				
6	2	0	2	Mark all rules
7	2	0	2	Mark all rules
(R4) *IF (h<177, AC<107,WG<5.15) THEN (GA<25.2)* (Supp=0.001,Conf=0.08,Spec=0.9,Sens=0.5,Lift=22.4,UN=0.9)				
(R5) *IF (h<183, WG<2.68) THEN (GA<17.5)* (Supp=0.001,Conf=0.33,Spec=0.88,Sens=0.1,Lift=179,UN=0.88)				
8	5	3	2	-

Analyzing the number of discovered rules reported in Table 4, one can see that by marking as uninteresting the rules satisfying a given pattern (in our case the pattern was defined by the form of the consequent term) the search was oriented towards rules not satisfying that pattern. On the other hand, as a consequence of the optimization criterion choice, some rules with very small support but a high lift value were discovered. Such rules correspond to exceptional cases which do not provide a general knowledge, but can be of interest for the medical doctors.

6. CONCLUSIONS AND FURTHER WORK

The strategy we proposed to allow the user to influence the process of rules' discovery is only a first step in developing an interactive system aimed at supporting medical doctors in exploring the data and extracting new, possibly unexpected knowledge. Despite its simplicity, the strategy allows the guidance of the searching process towards regions of higher interest.

The results we obtained with the obstetrical data were not as relevant as we might have expected due to multiple factors: on the one hand, the particularities of the data (many missing and erroneous values); on the other hand, the limitations of the evolutionary strategy itself. Using numerical values for the continuous attributes in order to avoid a preliminary discretization was appealing, but this led to a very large searching space and to the discovery of rules which were not easy to interpret. Using fuzzy variable instead of crisp ones could improve the quality of the final rules, especially in case of medical data.

Taking all these into account, we plan to adapt the searching strategy for fuzzy rules. Moreover, when having more than two optimization criteria, the non-dominated selection strategy we used is not able to sustain the diversity of the Pareto front, so another problem to be addressed is improving the selection strategy of the MOEA in order to deal with a high number of optimization criteria.

7. ACKNOWLEDGMENTS

This work is supported by the grant 99-II CEEX 03 – INFOSOC 4091/31.07.2006 from the Romanian Ministry of Education and Research (http://www.maternqual.ro).

8. REFERENCES

[1] Alatas, B., Akin, E., and Karci, A. 2008. MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing 8 (1), 646-656.

[2] Araujo, D.L.A., Lopes, H.S., and Freitas., A.A. 2000. Rule discovery with a parallel genetic algorithm, Proc. 2000 Genetic and Evolutionary Computation, Las Vegas, NV, USA, 89-92.

[3] Carvalho, D.R., Freitas., A.A., and Ebecken, N.N. 2005. Evaluating the correlation between objective rule interestingness measures and real human interest, Proc. European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-2005). LNAI 3721, Springer, 453-461.

[4] Deb, K., Agrawal S., Pratab, A., and Meyarivan, T., 2000. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In M. Schoenauer, et al. (eds), Proceedings of PPSN, Lecture Notes in Computer Science, vol 1917, 849-858.

[5] Deb, K, Chauduri, S. 2005. I-EMO: An interactive evolutionary multi-objective optimization tool, KanGAL Report 2005003, http://www.iitk.ac.in/kangal (Last access 4th April 2008).

[6] Dehuri, S, Patnaik, S., Ghosh, A., and Mall., R. 2008. Application of elitist multi-objective genetic algorithm for classification rule generation. Applied Soft Computing 8(1), 477-487.

[7] Fidelis, M.V., Lopes., H.S., and Freitas, A.A. 2000. Discovering comprehensible classification rules with a genetic algorithm, Proceedings of CEC2000, La Jolla, CA, USA, 805-810.

[8] Freitas, A. 2002. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer Verlag, Berlin.

[9] Gopalan, J., Alhajj, R., and Barker, K. 2006. Discovering accurate and interesting classification rules using genetic algorithms. In S. F. Crone, S. Lessmann, R. Stahlbock (Eds.): Proceedings of the 2006 International Conference on Data Mining, DMIN 2006, Las Vegas, Nevada, USA, June 26-29, 2006. CSREA Press, 389-395.

[10] Ishibuchi, H. and Yamamoto, T. 2006. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. PPSN Workshop on Multiobjective Problem Solving from Nature (Reykjavik, Iceland, September 9-13), http://dbkgroup.org/knowles/MPSN3/ (Last access 4th April 2008).

[11] Mata, J., Alvarez, J.L., and Riquelme, J.C. 2002. An Evolutionary Algorithm to Discover Numeric Association Rules. In Proceedings of SAC 2002, 590-594,.

[12] Noda, E., Freitas, A.A., and Lopes, H.S. 1999. Discovering interesting prediction rules with a genetic algorithm, Congress on Evolutionary Computation (CEC-99), Washington D.C., USA, 1322-1329.

[13] Ohsaki, M., Abe, H., Tsumoto, S., Yokoi, H., and Yamaguchi., T. 2006. Proposal of medical KDD support user interface utilizing rule interestingness measures. In Proceedings of ICDMW'06, IEEE, 759-764.

[14] Shillabeer, A. and Roddick, J.F. 2005. Reconceptualising interestingness metrics for medical data mining. Australian Workshop on Health Data mining, http://acrc.unisa.edu.au/groups/health/hdw2005/Shillabeer.pdf (Last access 4th April 2008).