A Multicriteria Genetic Algorithm to analyze DNA microarray data

Mohammed Khabzaoui, Clarisse Dhaenens and El-Ghazali Talbi
LIFL – UMR CNRS 8022 – Bâtiment M3
Université des Sciences et Technologies de Lille
59655 Villeneuve d'Ascq Cedex – FRANCE
Email: {khabzaou,dhaenens,talbi}@lifl.fr

Abstract— **Knowledge discovery from DNA microarray data has become an important research area for biologists. Association rules is an important task of knowledge discovery that can be applied to the analysis of gene expression in order to identify patterns of genes and regulatory network. Association rules discovery may be modelled as an optimization problem. In this paper, we propose a multicriteria model for association rules problem and present a Genetic Algorithm designed to deal with association rules on DNA microarray data, in order to obtain associations between genes. Hence, we expose the main features of the proposed Genetic Algorithm. We emphasize on specificities for the association rule problem (encoding, mutation and crossover operators) and on its multicriteria aspects. Results are given for real datasets[1].**

I. Introduction

DNA Microarray experiments have a great interest for biologists, thanks to their ability to simultaneously measure the expression and interactions of thousands of genes [5]. Genes of similar function may yield similar expression patterns in microarray hybridization experiments. Although microarrays have been applied in many biological studies [7], [11], [26], [4], [23], [17], the analysis of the large volumes of data generated (large matrices of expression levels of genes under different experimental conditions) is not trivial and requires advanced knowledge discovery processes. There exists several kinds of representations available to express knowledge that can be extracted from microarray data.

Many data mining techniques have been proposed to analyze microarray data. Here, we propose to analyze those data through the association rule technique, in order to determine associations between some different regulated genes.

In this work, the association rule problem is modelled as a multicriteria combinatorial optimization problem. We propose to solve it using an evolutionary algorithm based on genetic algorithms. Therefore, specific mechanisms (mutation and crossover operators, elitism, ...) have been designed.

In this paper, we firstly present the microarray technology. As we want to analyze these data with association rules, the third part presents a multicriteria model for the association rule problem. The fourth part proposes an evolutionary algorithm to extract association rules in a multicriteria way. It describes the algorithm, the designed operators, and its multicriteria aspects. Finally, results are given on real datasets.

[1]This works has been partially supported by the Genhomme project (French Research ministery)

II. Microarray data

A. Microarray technology

The microarray technology is now widely used in many areas of biomedical research. It provides access to expression levels of thousands of genes at once in order to identify co-expressed genes, relationships between genes, patterns of gene activity, changes in gene activity of some medical treatments... This technology consists in spotting probes that match a single gene (eg. complementary DNA of known genes) onto small slides of glass. Then DNA (in particular cDNA) is extracted from cells in a reference sample and a test sample, and labeled with different fluorescent dyes. These labeled cDNA are mixed and hybridized with the probes. Hence measuring the fluorescence (by scanning, for example) allows to calculate the relative abundance of cDNA of each gene and then to calculate their expression level.

There are two main variants of the DNA microarray technology: Synteni/Stanford chips and Affymetrixchips. They differ in:

- How DNA sequences are put down (spotting / photolithography).
- The length of DNA sequences (Complete sequences or a serie of fragments).

B. Preprocessing microarray data

Analyzing DNA microarray data requires a preprocessing phase [5].

1) Gene expression level: The differential gene expression is calculated by dividing the intensity of the gene in the sample under study by its intensity level in the control. This intensity ratio has a highly asymmetric distribution. To avoid this problem, a log_2-transformation is usually used to make a normal like distribution.

There are many sources of variations of measures in microarray experiments (variations in cells or individuals, mRNA extraction, isolation, hybridization condition, optical measurement, scanner noise etc...). Normalization can remove

0-7803-8515-2/04/$20.00 ©2004 IEEE

such variations. Many techniques for normalization aim to make the data more normally distributed (Log-transformation, per chip and per gene). This is an important issue to be able to analyze data [5].

2) Analyzing microarray data: Many data mining techniques have been proposed to analyze microarray data. Cluster analysis and classification techniques have been proposed to identify genes expression profiles. Eisen et al. [12] grouped genes sharing similar types of behavior over experiments into clusters. Kurra et al. [22] described a classification method to discriminate two types of leukemia using heuristic feature selection and a certain variant of perceptron-based classification method that separates these two classes of leukemia. Friedman et al. [15] used Bayesian networks to analyze gene expression data and to discover and describe interactions between genes. Recently, Deb and Reddy proposed a multicriteria Genetic Algorithm to obtain a reliable classification in cancer data [10].

Clustering methods allow biologists to design experiments helping them to understand further the relationships among the genes. However, they do not provide deep insights into specific relationships among the genes to be able to understand underlying biological processes in the cell.

Here, we propose to deal with such data using association rules. This is a more general model that allows to find associations between subsets of genes. Moreover, relations obtained are more precise thanks to the notions of condition and prediction. Finally association rules allow to deal with very large data sets.

3) Data and missing values: We are interested in finding genes that show significant changes between two groups of patients. Hence, we propose to remove genes that are equally expressed in the test and reference samples over the experiments.
Moreover, there may have numerous missing values in microarray data due to the empty spots or because the background intensity is higher than the spot intensity. Two ways are commonly used for the treatment of missing values: they may be replaced by estimated values (Median for example), or corresponding instances are deleted. Our approach is different. As we want to deal with these data through association rules, we propose that all genes that have missing values are keeped without modifications. Then, when these missing values concern the attributes of a rule, these genes are excluded from the computation of the quality of this rule.

III. Multicriteria association rules

We have shown the interest of using association rules to analyze microarray data. In this section, we present the multicriteria model we propose to deal with this rule discovery problem. Hence, we first present the classical association rule model for DNA microarray data. Then, the main question, to be able to use an optimization approach, is to define the optimization criteria. Therefore, we present a survey of the most famous criteria used to evaluate quality of rules. Finaly we expose results of a statistical analysis of these criteria which shows that no criterion is universal and that several criteria have to be jointly used in order to evaluate rules in a complete manner.

A. Association rules

The association rule problem was first formulated in [1] and is also called the market-basket problem. The problem is the following: given a set of items and a large collection of transactions which are sets (baskets) of items, the task is to find relationships between associations contents of various items within those baskets.

An association rule is an expression of the form: *IF* $cond_1$ *AND* $cond_2$ ... *AND* $cond_m$ *THEN pred.* This kind of rules contains two parts: the IF part which is called the rule condition (C) and the THEN part which is called the rule prediction (P). The rule condition contains a conjunction of m conditions about values of predictor attributes. Hence, both the C and the P parts are conjunctions of terms.

B. Association rules for DNA microarray data

Microarray data (gene expression levels) may be presented in a very similar way than the Market Basket data format. There are two ways to present DNA microarray data [21]:

- Gene table: Genes constitute the rows whereas treatments, to which the genes were exposed, are the columns. Clustering and classification may be applied to this table (clusters of genes).
- Treatment table: Experiments or treatments constitute the rows in the data-table whereas genes form the columns. In this case, association rules can be applied to the analysis of gene expressions in order to identify a set of regulated genes (associations between genes).

In our study we will consider data in the treatment table form and will look for rules combining genes, where a term can be in the form *<gene = value>*. Value belongs to the discretized gene expression level. An example of a rule could be : *IF* ($gene_{12}$ = *over_expressed) AND* ($gene_{504}$ = *under_expressed) THEN* ($gene_{8734}$ = *over_expressed).*

C. Quality criteria of a rule

In order to solve association rule discovery problem as a combinatorial optimization problem, the optimization criterion has to be defined. A lot of measures exist for estimating the quality of association rules. For an overview, readers can refer to Freitas [14], Tan et al. [27] or Hilderman et al. [18].
In this section, we survey and describe widely used measures proposed by different scientific communities. We study eleven (the most commonly used measures) criteria which have been introduced by statistics, probabilities, information theory and datamining communities. All these measures are presented with their formula in table I.

Formulas are given for a set of N instances, where $|C|$ represents the number of instances satisfying the C part of

TABLE I

QUALITY CRITERIA FOR ASSOCIATION RULES.

Measure	Formula
S	$\frac{\lvert CandP\rvert}{N}$
Cf	$\frac{\lvert CandP\rvert}{\lvert C\rvert}$
L	$\frac{\lvert CandP\rvert+1}{\lvert C\rvert+2}$
I	$\frac{N*\lvert CandP\rvert}{\lvert C\rvert*\lvert P\rvert}$
V	$\frac{\lvert C\rvert*\lvert P\rvert}{N*\lvert CandP\rvert}$
R	$\frac{\lvert CandP\rvert-\lvert CandP\rvert}{\lvert P\rvert}$
ζ	$\frac{\lvert CandP\rvert}{\lvert C\rvert+\lvert P\rvert-\lvert CandP\rvert}$
ϕ	$\frac{(\lvert CandP\rvert*\lvert CandP\rvert-\lvert CandP\rvert*\lvert CandP\rvert)^2}{\lvert C\rvert*\lvert P\rvert*\lvert C\rvert*\lvert P\rvert}$
IS	$\frac{\lvert CandP\rvert}{\sqrt{\lvert C\rvert*\lvert P\rvert}}$
J	$\frac{\lvert P\rvert}{N} * [\frac{\lvert CandP\rvert}{\lvert P\rvert} \log(\frac{N*\lvert CandP\rvert}{\lvert C\rvert*\lvert P\rvert}) + (1-\frac{\lvert CandP\rvert}{\lvert P\rvert}) \log(\frac{1-\frac{\lvert CandP\rvert}{\lvert P\rvert}}{1-\frac{\lvert C\rvert}{N}})]$
PS	$\frac{\lvert CandP\rvert}{N} - \frac{\lvert C\rvert}{N} * \frac{\lvert P\rvert}{N}$

the rule, $|P|$ the number of instances satisfying the P part of the rule and $|C\ and\ P|$ the number of instances satisfying simultaneously the C and the P parts of the rule.

Support (S): It is the classical measure of association rule. It enables to measure rule frequency in the database. It is the percentage of transactions containing, both the C part and the P part, in the database.

Confidence (Cf) **and Laplace** (L): The Confidence measures the validity of a rule. It is the conditional probability of P given C. Laplace is another measure of validity. It is used to measure precision of classification rules [8].

Interest (I): The Interest measures the dependency while privileging rare patterns in the region of weak support.

Conviction (V): The Interest has a symmetric behavior. To avoid this problem, Brin et al. [6] have proposed a new criterion "the conviction". It measures the quality of a rule as an implication.

Surprise (R): It is used to measure the affirmation. It enables to search surprising rules.

Jaccard (ζ): This measure of similarity is usually used to compute the distance between two text words.

Phi-coefficient (ϕ): It is a dependency measure which is derived from the Chi-square test. This measure is similar to the correlation coefficient in its interpretation. The phi-coefficient measures the degree of association between two binary variables in statistic.

Cosine (IS): It is derived from the statistic correlation and is very interesting in the region of weak Support and strong Interest. It is the geometric average of the confidence of two rules.

J-measure (Jm): Smyth and Goodman [25] have proposed the J-measure, which estimates the degree of interest of a rule and combines support and confidence. It is used in optimization [28], [2].

Piatetsky-Shapiro (PS): It is a measure of dependency proposed by Piatetsky Shapiro [24]. It favors rules which predict minority classes.

D. Analysis of these criteria

We aim to focus on the correlations between criteria in order to determine a coherent set of complementary criteria (group of uncorrelated criteria). To study relations between criteria, we enumerate all existing rules for a given problem of association rules. Then, we measure each rule according to the eleven criteria and generate a table where a row represents an association rule and each column represents the quality of this rule according to one criterion. We study this data table thanks to the Principal Components Analysis (PCA).

To find strong correlations between criteria we have to find correlations between columns of this matrix. These correlations are summarized by Table II (matrix of linear correlations between the eleven criteria).

We can remark a strong correlation between Cosine and Jaccard (0.98), between J-Measure and Piatetsky (0.93), between Confidence and Laplace (very similar measures, 0.99), which are candidate to form three different groups. Support is correlated with Jaccard (0.87), then it can be attached to the first group. In the same manner, Phi-Coefficient and Conviction can be attached to the second group thanks to their correlations with the Jmeasure. Interest and Surprise criteria seem to be atypical criteria and form two other groups. These five groups are summarized in Table III.

Group 1	Jaccard, Cosine, Support
Group 2	Laplace, Confidence
Group 3	Phi-Coefficient, Conviction, J-measure, Piatetsky
Group 4	Interest
Group 5	Surprise

TABLE III

GROUPS OF CRITERIA.

An analysis in PCA (Principal Components Analysis) was made on this correlation matrix. Figure 1 shows the circle of correlation on the two main axis (with 75.44 % of inertia) and confirms our previous affirmations (proposed groups of criteria may be distinguished on Figure 1) and proves the significance of the first factorial axis. The exposed statistical analysis of the different criteria allows us to group them into five groups, where each group is composed of correlated criteria. This study has been lead for several databases (UCI, Microarray data, ...) and gives similar results.

Here we propose, to deal with the association rule problem, to model it as a multicriteria problem. Therefore, we choose, for each group, one criterion, and obtain five complementary criteria that allow to evaluate rules in a complete way taking into account specificities of several criteria.

	S	Cf	I	V	R	ζ	ϕ	IS	J	PS	L
S	1.00										
Cf	0.62	1.00									
I	-0.09	0.20	1.00								
V	0.27	0.56	0.47	1.00							
R	0.17	0.48	0.07	0.17	1.00						
ζ	**0.87**	0.62	0.32	0.55	0.20	1.00					
ϕ	0.38	0.50	0.62	**0.81**	0.26	0.76	1.00				
IS	**0.86**	0.68	0.34	0.56	0.19	**0.98**	0.76	1.00			
J	0.34	0.50	0.40	**0.84**	0.15	0.64	**0.89**	0.62	1.00		
PS	0.29	0.49	0.25	0.71	0.15	0.51	0.75	0.51	**0.93**	1.00	
L	0.63	**0.99**	0.18	0.54	0.53	0.61	0.49	0.67	0.50	0.51	1.00

TABLE II
MATRIX OF LINEAR CORRELATIONS.

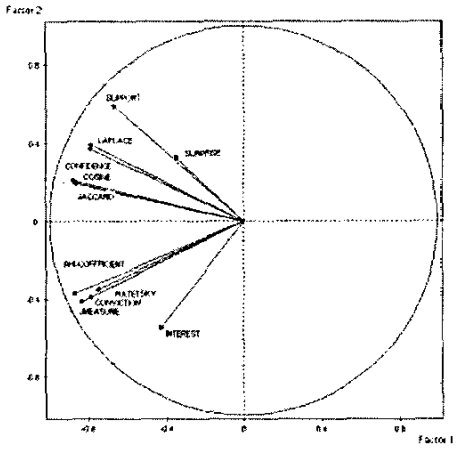

Fig. 1. Circle of correlation.

IV. A MULTICRITERIA GENETIC ALGORITHM

In order to deal with the multicriteria association rule problem we develop a specific Genetic Algorithm (GA) [16], as GAs have been widely used to solve multicriteria optimization problems and have been applied with success to monocriterion rule mining problems [20]. We adopt a "A posteriori" approach while looking for all the solutions of best compromise between criteria. We expose the main features of the proposed Genetic Algorithm scheme for association rules problem (encoding, mutation and crossover operators) and its multicriteria aspects (archive of Pareto solutions, ranking, management of the population).

A. A general scheme

Genetic algorithms are inspired by Darwin's theory of evolution. The algorithm starts with a set of randomly generated solutions (population). Then, solutions are selected, according to their fitness quality to form new solutions (offspring). In a multicriteria scheme, best solutions encountered over generations are filed into a secondary population called the "Pareto Archive".

Figure 2 presents the Multicriteria genetic algorithm scheme. We choose in this model the five following criteria (one of each group): Support, Jmeasure, Interest, Surprise and Confidence.

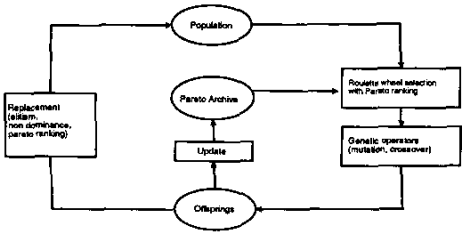

Fig. 2. A multicriteria genetic algorithm.

B. Operators for association rules

Crossover and mutation operators are two basic operators of genetic algorithms.

Crossover: The crossover mixes the features of two rules by the combination of their attributes. The proposed crossover operator has two versions:

- **Crossover by value mutation** If two rules X and Y have one or several commonattribute(s) in their C parts, one common attribute is randomly selected. The value of the selected attribute in X is exchanged with its counterpart in Y(see Figure 3).
- **Crossover by insertion** Conversely, if X and Y have no commonattribute, one term is randomly selected in the C part of X and inserted in Y with a probability inversely proportional to the length of Y. The similar operation is performed to insert one term of Y in X(see Figure 4).

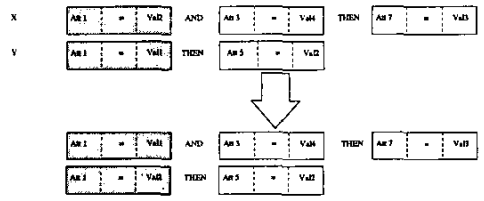

Fig. 3. Crossover by value mutation.

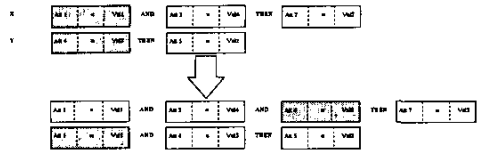

Fig. 4. Crossover by insertion of attributes.

Mutation: First, two mutation operators were implemented. The mutation operation called "Value mutation" replaces an attribute value by a randomly chosen one (see Figure 5). The second one, called "Attribute mutation", replaces an attribute by another. The value of this attribute is randomly chosen in its domain (see Figure 6).

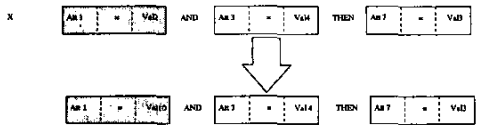

Fig. 5. Value mutation.

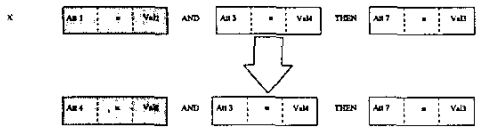

Fig. 6. Attribute mutation.

Adaptive mutation rate: Then, two other mutation operators are proposed. The insertion operator that adds a randomly chosen attribute in the rule, and the delete operator that removes an attribute of the rule (if the number of attributes is greater or equal to 3). Hence setting the probabilities of appliance of these four mutation operators may be difficult. Moreover, the interesting operator at a given time of the search is not always the same. Therefore, we propose to set these probabilities of appliance in an adaptively way where the more efficient an operator is, the more it will be used, as proposed in [19].

C. Multicriteria aspects

Dealing with a multicriteria problem requires adaptation of optimization methods. The main difference, compared to monocriterion optimization problem is that in multiobjective problem, there is not a single solution for which all criteria are optimal but, a set of solutions for which there are no other solutions better for all the criteria. These solutions are called Pareto-optimal. The notion of Pareto-optimality is defined in terms of dominance. Let's consider a multicriteria minimization problem with k criteria. A solution $x = (x_1, x_2, \ldots, x_k)$ is said dominating another solution $y = (y_1, y_2, \ldots, y_k)$ if $\forall i$, $x_i \leq y_i$ and $\exists i \ / \ x_i < y_i$. A solution x is a member of the Pareto-set, or said to be non-dominated, if there is no other solution w such that w dominates x.

We propose to use a genetic algorithm to find all the Pareto-optimal solutions (solutions of best compromise) which are all interesting potential rules. They are located on a boundary, known as the Pareto-front. We would like the solutions to cover the Pareto-front as well as possible, to obtain a good representation of this front. This approach offers multiple solutions to the decision maker that can select the solution that is best suited according to additional criteria, without requiring additional searches.

In order to deal with multicriteria optimization problems, different mechanisms have to be used. For example, the notion of dominance has to be defined (to be able to compare solutions between them) and population management has to be carefully studied.

Selection operator: We use the classical roulette selection based on the ranking notion. The probability of selection of a solution is proportional to its rank. We use two ranking methods:

1 - Pareto ranking[13]: The rank of a solution corresponds to the number of solutions, in the current population, by which it is dominated (see Figure 7 for a minimization problem).

2- Non-Dominated Sorting GA (NSGA)[9]: This method assigns ranks to solutions by first finding the set of non-dominated solutions in the current population. Those solutions are removed from the population and assigned rank 1. As these solutions are removed, a new so-called front of non-dominated solutions is now present in the remainder of the original population. This second front is extracted and assigned rank 2. This procedure is repeated until no more solutions are present in the population (see Figure 8 for a minimization problem).

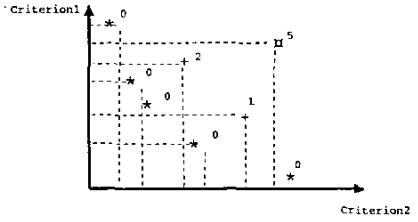

Fig. 7. Pareto ranking.

Replacement operator: We use the elitist non dominated sorting replacement. The worst ranked solutions are replaced by dominating solutions (if there is any) generated by

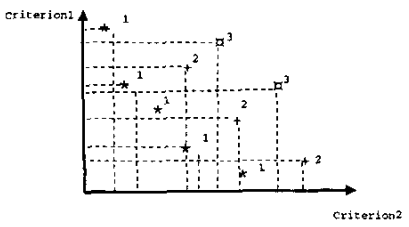

Fig. 8. NSGA ranking

mutation and crossover operators (offspring). The size of the population remains unchanged.

Archive: Non dominated association rules are archived into a secondary population called the "Pareto Archive" in order to keep track of them. It consists in archiving all the Pareto association rules encountered over generations. This archive has to be updated each time a solution is added.

Elitism: The Pareto solutions (best solutions) are not only stored permanently, they also take part in the selection and may participate to the reproduction.

V. Results

In order to evaluate the algorithm, we test it on two microarray databases:

- a confidential database containing 22376 human genes for 45 Affymetrix chips (DB1).
- a public database "MIPS Yeast Genome Database" containing 2467 genes for 79 chips (YeastDB).

The evaluation of the algorithm has been done is two phases. First we study the different operators proposed and evaluate their efficiency on microarray datasets. Then in a second phase, we analyze rules obtained on a real dataset.

A. Analysis of operators

In order to compare two Pareto fronts, we use the **contribution** indicator [3]. It quantifies the domination between two sets of non-dominated solutions $(Front1, Front2)$ through the ratio of non dominated solutions produced by each of these two sets $(Cont(Front1/Front2) \in [0,1]$ and $Cont(Front1/Front2) + Cont(Front2/Front1) = 1)$.

Contribution	Pareto without elitism	NSGA without elitism
Pareto with elitism	0.64	
NSGA with elitism		0.75

TABLE IV
CONTRIBUTION OF ELITISM (DB1).

Table IV indicates the contribution of the use of **elitism** during the selection phase. It shows, as its contribution is greater than 0.5, that elitism improves the selection with the two studied operators (Pareto ranking and NSGA). So it seems important to use elitism while selecting individuals.

Contribution	Pareto with elitism	NSGA with elitism
Pareto with elitism		0.54
NSGA with elitism	0.46	

TABLE V
CONTRIBUTION: PARETO RANKING / NSGA SELECTION (DB1).

In order to know which **ranking method** to use, we also compare the Pareto ranking with the NSGA. Table V indicates that their efficiencies are on the same range with a small preference for Pareto ranking. So, this will be the one we will use for the rest of our experiments.

Two versions concerning the mutation operators have been implemented. The first one, called the non adaptive, only uses the attribute mutation and the value mutation. The second one, called the adaptive, uses the fourth mutation operators in an adaptive way. In order to compare these two versions, we ran 10 executions on both databases. Table VI indicates, for the two databases, first the number of solutions of fronts obtained thanks to the both versions and the contribution of the adaptive version. It shows that the adaptive version produces more solutions that constitute a better final front than the one obtained with non adaptive version.

	DB1	YeastDB
Non adaptive	9 sol.	19 sol.
Adaptive	12 sol.	31 sol.
Contribution adaptive/non adaptive	0.54	0.71

TABLE VI
ADAPTIVE VS NON ADAPTIVE.

This result may be explained thanks to Figures 9 and 10 which show the evolution of the contribution, between the Pareto front at the end of a generation and the Pareto front at the beginning of the generation, over generations. When the contribution increases, it indicates that at the given generation an important improvement of the Pareto front has been realized. This two figures show that with the adaptive version of the algorithm, improvements occur longer than for the non adaptive version. This yields to better find solutions.

B. Use on real data

The second part of our experiments deals with analyzing results produced for microarray data. For the following tests, the adaptive version with Pareto ranking and elitism is used with the following parameters:

- Size of the population: DB1 = 200, YeastDB = 1000

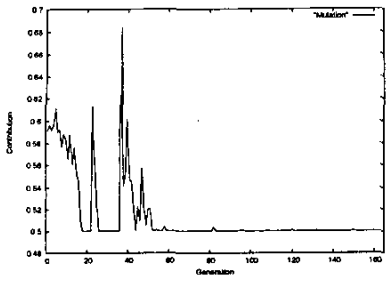

Fig. 9. Contribution over generations with non adaptive version.

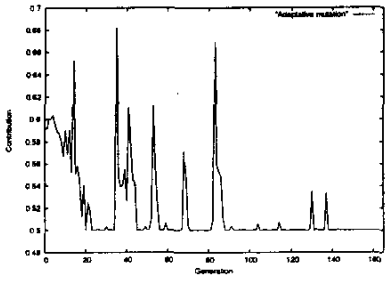

Fig. 10. Contribution over generations with adaptive version.

- Pareto archive selection (elitism): 0.5
- Crossover probability: 0.8
- Mutation probability: 0.4
- Number of generations: 200

In both databases, gene expressions have been discretized thanks to the affymetrix process and may take five values: Increase (I), Marginal Increase (MI), when the gene is over expressed, Decrease (D) and Marginal Decrease (MD) when it is under-expressed and No Change (NC), when the difference of expression is not significative.

For a first study on DB1, a set of 514 genes that show an interesting differential expression over the set of experiments (filtered on the number of No Changes), have been selected. On the contrary, the whole Yeast DB has been used.

In order to validate the multicriteria approach, some 2D projections of the pareto front are given for some pairs of criteria on Figures 11 and 12. On these figures only solutions of the pareto front of the two considered criteria are presented. We can conclude that those pairs of criteria are really complementary as solutions good for one criteria are not very good for the second one. Moreover, the fronts have a large number of solutions that represent different levels of compromise.

Table VII shows some Pareto solutions found for DB1 using the multicriteria model and table VIII indicates their values for the five criteria (Support, Jmeasure, Interest, Surprise and Confidence).

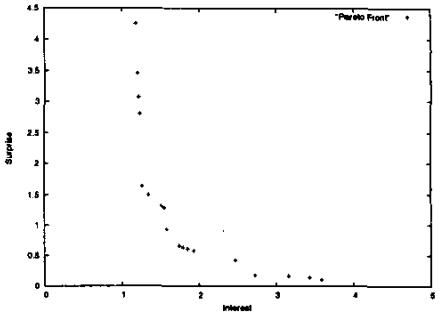

Fig. 11. Pareto Front (Surprise / Interest) - YeastDB.

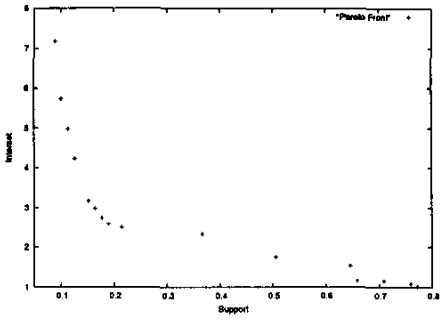

Fig. 12. Pareto Front (Interest / Support) - YeastDB.

TableVIII shows one more time the interest of using the proposed multicriteria model as good solutions for one criteria are not always interesting for another one.

Rules	Description
R1	if ((122=I) and (372=NC) and (499=NC) and (435=NC)) then (222=I)
R2	if ((436=NC) and (63=I) and (487=NC) and (332=NC) and (210=I) and (374=NC)) then (219=I)
R3	if ((161=I) and (229=I) and (118=D) and (503=NC) and (311=NC)) then (39=D)
R4	if ((238=I) and (226=I) and (426=NC))) then (64=I)
R5	if((146=I) and (318=NC) and (499=NC) and (435=NC) and (457=NC) and (479=NC) and (367=NC) and (457=NC)) then (222=I)

TABLE VII
DESCRIPTION OF SOME PARETO SOLUTIONS OBTAINED (DB1).

For the Yeast Database, results are similar. Here are some genes participating to rules of the Pareto set: $IDP1$, $ENO2$, $ALG11$, $PRS5$, $CAF20$, $MRPL20$, $PET9$. These genes and relations existing between them form assumptions that must be validated by biologists.

Rules	S	Jm	I	R	Cf
R1	0.133	0.248	6.428	0.857	1.000
R2	0.155	0.268	5.625	0.857	0.875
R3	0.244	0.268	3.000	0.733	1.000
R4	0.244	0.268	3.000	0.733	1.000
R5	0.688	0.153	1.250	0.838	0.861

TABLE VIII
QUALITY OF SOME PARETO SOLUTIONS OBTAINED (DB1).

VI. CONCLUSION

In this work, we propose a multicriteria model to extract association rules from DNA Microarray Data. We propose to solve it using a multicriteria genetic algorithm that has been developed, in our case, thanks to the EO (Evolving Objects) platform `http://eodev.sourceforge.net` that allows to easily develop evolutionary algorithms once you have designed its specificities. Hence, we exposed the main features of the proposed Genetic Algorithm scheme for association rule problem and its multicriteria aspects. Different operators and mechanisms were compared in order to determine the most suitable scheme. This algorithm has been used to deal with microarray data. It has shown its ability to produce rules describing relations between genes, its robustness and its efficiency while producing similar rules from one execution to another.

The method proposed in this article is able to deal with any data coming from DNA microarray experiments. It now has to be used on other datasets. Association rules problem has been presented here as a high complexity multicriteria combinatorial problem. In order to be able to deal with a larger number of genes (without any filtering phase, for example), a perspective could be the development of a parallel genetic algorithm.

Moreover, the multicriteria model lead to the proposition of a large number of rules that have to be studied. Therefore we devellop ARV (`http://www.lifl.fr/~jourdan/download/arv.html`) a visualizing tool to help the selection of rules in a multicriteria context.

REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th Intl. Conference on Very Large Databases,Santiago, Chile*, Sept 1994.

[2] D.L.A. Araujo, H.S. Lopes, and A.A. Freitas. A Parallel Genetic Algorithm for Rule Discovery in Large Databases. In *Proc. 1999 IEEE Systems, Man and Cybernetics Conf.*, volume III, pages 940–945, Tokyo, Japan, October 1999.

[3] M. Basseur, F. Seynhaeve, and E-G. Talbi. Design of multi-objective evolutionary algorithms: Application to the flow-shop scheduling problem. *Congress on Evolutionary Computation (CEC'02),Honolulu, USA*, pages 1151–1156, 2002.

[4] MA. Behr, MA. Wilson, and WP. Gill WP. Comparative genomics of bcg vaccines by whole-genome dna microarray. *Science*, 1520-3:284, 1999.

[5] D.P. Berrar, W. Dubitzky, and M. Granzow. *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, 2003.

[6] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings ACM SIGMOD, USA*, pages 255–264, 1997.

[7] PO. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature Genetics Supplement*, 21:33–37, 1999.

[8] P. Clark and R. Boswell. Rule induction with CN2: Some recent improvements. In *Proc. Fifth European Working Session on Learning*, pages 151–163, Berlin, 1991. Springer.

[9] K. Deb, S. Agrawal, A. Pratab, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *Proceedings of the Parallel Problem Solving from Nature VI Conference*. Springer. Lecture Notes in Computer Science No. 1917, 2000.

[10] K. Deb and A. R. Reddy. Reliable classification of two-class cancer data using evolutionary algorithms. *BioSystems*, 72:111–129, 2003.

[11] DJ. Duggan, M. Bittner, Y. Chen, P. Meltzer, and JM. Trent. Expression profiling using cdna microarrays. *Nature Genetics Supplement*, 21:10–14, 1999.

[12] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of National Academy of Sciences,95(25):14863-8*, 1998.

[13] C. M. Fonseca and P. J. Fleming. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3:1:1–16, 1995.

[14] A. Freitas. On rule interestingness measures. *Knowledge-Based Systems journal*, 1999.

[15] N. Friedman, M. Linial, I.Nachman, and D. Peér. Using bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conferenceon on Computational Molecular Biology (RECOMB-2000),*, 2000.

[16] D. E. Goldberg. *Genetic Algorithms - in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company, 1989.

[17] JG. Hacia, LC. Brody, MS. Chee, SP. Fodor, and FS. Collins. Detection of heterozygous mutations in brca1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Med*, 441-7:14, 1996.

[18] R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey, technical report cs 99-04. Technical report, Department of Computer Science,University of Regina, October 1999.

[19] T.P. Hong, H. Wang, and W. Chen. Simultaneously applying multipke mutation operators in genetic algorithms. *journal of heuristics*, 6:439–455, 2000.

[20] L. Jourdan, C. Dhaenens, and E.G. Talbi. Rules extraction in linkage desequilibrium mapping with an adaptive genetic algorithm. In *Proceedings of the European Conference on Computational Biology (ECCB'03)*, pages 29–31, 2003.

[21] P. Kotala, P. Zhou, S. Mudivarthy, W. Perrizo, and E. Deckard. Gene expression profiling of dna microarray data using peano count trees. In *Online Proceedings of the First Virtual Conference on Genomics and Bioinformatics. URL: http://midas-10.cs.ndsu.nodak.edu/bio/.*, October 2001.

[22] G. Kurra, W. Niu, and R. Bhatnagar. Mining microarray expression data for classifier gene-cores. In *Proceedings of BIOKDD 01*, 2001.

[23] MJ. Marton, JL. DeRisi, and HA. Bennett. Drug target validation and identification of secondary drug target effects using dna microarrays. *Nat Genet*, 1293-301:4, 1998.

[24] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases, AAAI/MIT Press*, pages 229–248, 1991.

[25] P. Smyth and R. M. Goodman. *Knowledge Discovery in Databases*, chapter Rule Induction Using Information Theory, pages 159–176. Piatetsky-Shapiro G. and Frawley J, 1991.

[26] PT. Spellman, G. Sherlock, and MQ. Zhang. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9:3273–97, 1998.

[27] P-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eight ACM SIGKDD conference, Edmonton, Canada*, 2002.

[28] K. Wang, S. H. W. Tay, and B. Liu. Interestingness-based interval merger for numeric association rules. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro, editors, *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 121–128. AAAI Press, 27–31 1998. New York, USA.