EGEA : A New Hybrid Approach Towards Extracting Reduced Generic Association Rule Set (Application to AML Blood Cancer Therapy)

M.A. Esseghir, G. Gasmi, S. Ben Yahia, and Y. Slimani

Département des Sciences de l'Informatique
Faculté des Sciences de Tunis
Campus Universitaire, 1060 Tunis, Tunisie
mohamedemir@gawab.com

Abstract. To avoid obtaining an unmanageable highly sized association rule sets– compounded with their low precision– that often make the perusal of knowledge ineffective, the extraction and exploitation of compact and informative generic basis of association rules is a becoming a must. Moreover, they provide a powerful verification technique for hampering gene mis-annotating or badly clustering in the Unigene library. However, extracted generic basis is still oversized and their exploitation is impractical. Thus, providing critical nuggets of extra-valued knowledge is a compellingly addressable issue. To tackle such a drawback, we propose in this paper a novel approach, called EGEA (Evolutionary Gene Extraction Approach). Such approach aims to considerably reduce the quantity of knowledge, extracted from a gene expression dataset, presented to an expert. Thus, we use a genetic algorithm to select the more predictive set of genes related to patient situations. Once, the relevant attributes (genes) have been selected, they serve as an input for a second approach stage, i.e., extracting generic association rules from this reduced set of genes. The notably decrease of the generic association rule cardinality, extracted from the selected gene set, permits to improve the quality of knowledge exploitation. Carried out experiments on a benchmark dataset pointed out that among this set, there are genes which are previously unknown prognosis-associated genes. This may serve as molecular targets for new therapeutic strategies to repress the relapse of pediatric acute myeloid leukemia (AML).

Keywords: Generic association rules, Genetic Algorithms, Neural networks, Frequent Closed itemset algorithms, Bioinformatics.

1 Introduction

High-throughput sequencing and functional genomic technologies provided to the scientific community a human genome sequence and have enabled large-scale genotyping and gene expression profiling of human populations [1]. Biological databases contain heterogeneous information such as annotated genomic

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2006, LNCS 4081, pp. 491–502, 2006.

sequence information, results of microarray experiments, molecular structures and properties of proteins, etc. In addition, an increasing number of databases from the medical domain, containing medical records and valuable information on diseases and phenotypes, become available. Data Mining techniques and/or tools, aiming to go further beyond the top of the Iceberg, delve and efficiently discover valuable, non-obvious information from large microarray databases (*e.g.*, information about diseases and their relation to sub-cellular processes). Microarrays provide a prolific, "exciting" and challenging contexts for the application of data mining techniques. For recent overviews, please refer to recently edited books respectively by Wang *et al.* [1] and Chen [2].

In this respect, extracting generic basis of association rules seems to be an efficient approach for providing extra-added value knowledge for biologists. In this case, we expect that a biologist may not only discover synexpression groups but may also identify correlations between a group of genes and a particular cell type. However, the unmanageably large association rule sets, even though generic association rule set size is known to be compact, compounded with their low precision often make the perusal of knowledge ineffective, their exploitation time-consuming, and frustrating for the user.

In this paper, and aiming to tackle this highly important topic, we propose a novel approach towards reducing "shrewdly" and informatively the amount of knowledge to be presented to a user, we propose an hybrid approach showing the potential benefits from the synergy of genetic algorithms and association rule extraction. Thus, we used a genetic algorithm to select the more predictive set of genes related to the patient situation. Then, we extract generic association rules from this reduced set of genes. The notably decrease of the generic association rules, extracted from the selected genes, permits to ameliorate the quality of knowledge exploitation.

Experiments were carried out on a dataset of the AFFIMETRIX GENECHIP Human Genome U95Av2 oligonucleotide microarray (Affymetrix, Santa Clara, CA) that contains `12 566` probe sets. This dataset contains Analysis of mononuclear cells from 54 chemotherapy treated patients less than 15 years of age with acute myeloid leukemia (AML). Mononuclear cells taken from peripheral blood or bone marrow. Treatment results describing patient situation associated with *complete remission* and *relapse* with resistant disease are also reported. After the chemotherapy treatment, most patients with Acute Myeloid Leukemia (AML) enter complete remission. However, some of them enter relapse with a resistant disease. Obtained results showed that Also, among this set, there are genes which are previously unknown prognosis-associated genes. This may serve as molecular targets for new therapeutic strategies to repress the relapse of pediatric AML.

The remainder of the paper is organized as follows. Section 2 details the proposed hybrid approach. The genetic algorithm applied for the selection of most predictive attributes is described. Section 3 presents the obtained results from the carried out experiments on the benchmark dataset. Section 4 concludes this paper and points out future perspectives.

2 Dimensionality Reduction: Selection of a Predictor Set of Genes

Applying classical association rule extraction framework to dense microarrays leads to an unmanageably highly sized association rule sets– compounded with their low precision– that often make the perusal of knowledge ineffective, their exploitation time-consuming, and frustrating for the user. Even though extracting and exploiting compact and informative generic basis of association rules can be an advisable remedy, a glance to their size can be nightmarish to the user (c.f, reported statistics in Experiments section). Another avenue to tackle the high dimensionality problem in gene expression datasets, is to assess and select one of the more discriminatory set of genes to the target. In fact, *feature selection* refers to the problem of selecting the more predictive and valuable attributes, in terms classification and class separability, correlated with a given output. Numerous studies have focused on the selection of relevant features, by discarding misleading and noisy ones [3]. Such studies, involving different techniques can be viewed under two families: *wrappers* and *filters*. Wrappers evaluate attributes by using accuracy estimates provided by the actual target learning algorithm. Alternatively, filters use general characteristics of the data to evaluate attributes and operate independently of any learning algorithm [4]. Indeed, an exhaustive search within the large set of feature combination is very consuming in term of computational time, since the search space of possible combination increases exponentially with the number of genes. An exhaustive search of all possible combinations of attributes is impractical, especially when the evaluation procedure for the generated solutions involves a learning algorithm. In this respect, the use of AI global search techniques, such as genetic algorithms (GA) seems to be very promising, since they have proven to be valuable in the exploration of large and/or complex problem spaces [5]. GA attempt to apply evolutionary techniques to the field of the problem solving notably in combinatorial optimization [6,7]. In fact, GA may be used to select the more predictive set of genes related to the target class (patient situation). Our genetic algorithm evolves a set of feasible solutions evaluated with an artificial neural network as a wrapper. Subsets of variables are assessed within the evaluation procedure according to their generalization accuracy in classification.

Believing that combining classifiers and boosting methods can lead to improvements in performance, in this paper, we propose, a new hybrid model whose the driving idea is the go towards assessing potential benefits form a synergy of two data mining techniques, namely feature selection by Artificial neural networks and GA and association rule extraction.

Figure 1 graphically sketches the model and shows that it is composed of two steps. On, the first stage selects the best set of inputs having a predictive relationship with the target class. Whereas, the second step consists in the generation of a compact set of generic association rules using the selected genes. It is noteworthy that the whole process, i.e., sequentially applying feature selection and rule extraction, is performed in an iterative process. From an iteration to another one, and acting towards a more reduced and guided search space, the

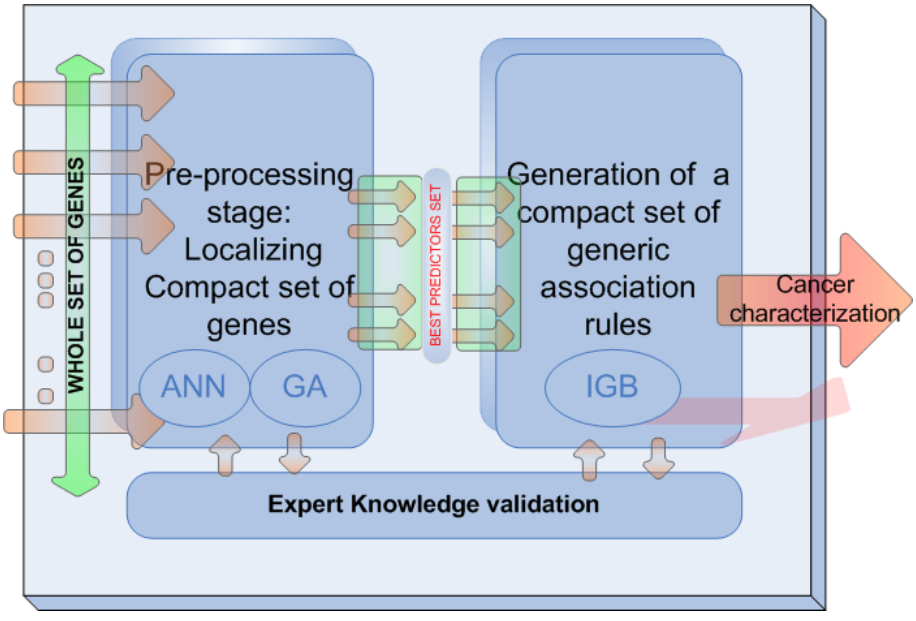

Fig. 1. The proposed hybrid model

system can be fed by biological apriori knowledge or given by experts or pointed out by generic association rules.

2.1 Preprocessing Stage: Dimensionality Reduction

In this section we look to the problem of building a representative set of relevant features. In fact Data reduction techniques was successfully applied in in numerous gene expression data analysis using as well wrapper as filters [8,9,10]. Narayanan *et al.* [11] have applied different data mining techniques, mainly based on neural network classifiers, to mine knowledge enfolded in microarrays data using also neural networks as a wrapper to tackle the high dimensional data.

Believing that feature selection methods, to use as well in the definition of a compact pattern representation as in mining knowledge with robust and interpretable methods, depends mainly both on wrapper accuracy -quality of the evaluation procedure- and on the search procedure applied. We decide to opt for a stochastic global search procedure to explore the search space of feasible subsets of relevant non-redundant features: *Genetic algorithms*. In addition, the wrapper consists of an artificial neural networks trained with the *backpropagation* learning algorithm [12]. The genetic algorithm, presented here, will be applied to select a subset of genes involved in the prediction of patient situation: *complete remission* or *relapse*. Each set of candidate solutions are evolved through a fixed number of generations. The pseudo-code for localizing compact predictive set of genes is provided by *Algorithm 1*. GA process is roughly summarized in what follows:

– *Representation*: Each generation consists of a set of candidate solutions represented using a binary encoding. Any possible solution to the problem is encoded as binary string of 12625 genes, where the code 1 means that the gene is selected to build an ANN and 0 when it is discarded (*c.f.*, Figure 2).

Fig. 2. Chromosome representation

– *Initialization*: An initial set of solutions is randomly generated. For each individual a number of genes are randomly selected by setting the corresponding bits equal to 1. Once the initial set is generated, the evaluation process starts. A fitness value is assigned to each chromosome. The first generation of solutions is derived by applying a tournament selection to the evaluated set.
– *Evaluation*: Two steps are required to evaluate each chromosome. First a neural network, with the selected genes in the chromosome as input, is built and partially trained. Next, the trained network is evaluated using the test set. The test set presents to the network a new data which is not trained with it. The chromosome evaluation assesses the predictive generalization ability of the neural network and consequently of the set of involved genes . Our fitness involves two evaluation criteria: the proportion of incorrectly classified instances and the mean square error on the test set.

$$fitness = (ICI + TMSE)/2 \tag{1}$$

Where ICI and $TMSE$ denotes respectively the proportion of incorrectly classified instances in the data set and the mean square error found on the test set.

2.2 Generic Association Rule Extraction

An association rule $R : X \Rightarrow Y - X$ is a relation between two frequent itemsets $X \subset Y$. X and $(Y - X)$ are called, respectively, premise and conclusion of the rule R. An association rule is valid whenever its strength metric, confidence(R)=$\frac{support(Y)}{support(X)}$, is greater than or equal to the minimal threshold of confidence *minconf*.

However, in practice, the number of valid association rules is very high. Indeed, this number was estimated to be equal to $2^{2\times l}$, where l is the length of the longest frequent itemset [13].

Consequently, the user can not interpret and exploit efficiently a such amount of knowledge. To palliate this problem, a solution consists in extracting a **reduced** subset of association rules, called *generic basis*. On the demand of the user, we have to be able to derive all the remaining association rules (*i.e.*., generic basis extraction should be done **without information loss**). For this reason, generic basis extraction have to fulfill the following requirements: [14]

Algorithm 1. Feature selection: Localizing compact set of genes

Input:
S: set of genes
N_i: initial population size
N: population size
t: tournament size
p_{mut}: mutation probability
p_{cross}: Crossover probability
it: number of training iterations
Maxgen: number of generations
h: number of hidden nodes
η: learning rate
m: momentum value
Output: S1: Best subset of gene predictors

```
Begin
    Population P0, P,Ptmp
    P0=P=Ptmp= ∅;
    P0=GenerateInitialPopulation(Ni)
    Evaluate (P0, η, m, h, it)
    P=Select(P0, N, t) //Applying a tournament to select N chromosomes from P0
    i=0
    While i < Maxgen do
        Ptmp=Select (P, N, t)
        Crossover(Ptmp, pcross)
        Mutate(Ptmp, pmut)
        Evaluate(Ptmp, η, m, h, it)
        Replace(Ptmp, P) //replacing solutions from P by newest ones from Ptmp using reverse tournament
        i=i+1
    S1=P.bestChromosome().extractGenes() // extracting selected chromosome genes
    Return(S1)
End
```

- "***Derivability***": An inference mechanism should be provided (*e.g.*, an axiomatic system). The axiomatic system has to be valid (*i.e.*, should forbid derivation of non valid rules) and complete (*i.e.*, should enable derivation of all valid rules).
- "***Informativeness***": The generic basis of association rules allows to retrieve exactly the support and confidence of the derived (redundant) association rules.

To extract a reliable number of association rules, we use the $\mathcal{IGB}$ (Informative Generic Basis) basis [14]. This choice is justified by:

1. **Conveying maximum of useful knowledge:** Association rules of the $\mathcal{IGB}$ basis convey the maximum of useful knowledge. Indeed, $\mathcal{IGB}$ is defined as follows:

Definition 1. *Let $\mathcal{FCI}_{\mathcal{K}}$ be the set of frequent closed itemsets* [1] *extracted from an extraction context $\mathcal{K}$. For each entry f in $\mathcal{FCI}_{\mathcal{K}}$, let MG_f be the set of its minimal generators* [2]*. The $\mathcal{IGB}$ generic basis is given by: $\mathcal{IGB} = \{R : g_s \Rightarrow (f\text{-}g_s) \mid f \in \mathcal{FCI}_{\mathcal{K}} \wedge f \neq \emptyset \wedge g_s \in MG_{f_1}, f_1 \in \mathcal{FCI}_{\mathcal{K}} \wedge f_1 \subseteq f \wedge confidence(R) \geq minconf \wedge \nexists\, g \,/\, g \subset g_s \wedge confidence(g \Rightarrow f\text{-}g) \geq minconf\}$.*

Thus, a generic association rule of $\mathcal{IGB}$ is based on the extraction of frequent closed itemsets from whose we generate minimal generic association rules, i.e., with minimal premise part and maximal conclusion part. It was shown that this type of association rules conveys the maximum of useful knowledge [15];

2. **Information lossless:** It was pointed out that the $\mathcal{IGB}$ basis is extracted without information loss [14];
3. **Compactness:** In [14] and by comparing obtained set cardinalities, we showed that $\mathcal{IGB}$ is by far more compact than the following:
   - The Non-Redundant association Rules $\mathcal{NRR}$ basis, defined by Zaki et *al.* [16,17];
   - The Generic Basis of Exact rules and the Transitive reduction of Generic Basis of Approximative rules ($\mathcal{GBE}$, $\mathcal{TGBA}$), defined by Bastide et *al.* [18].

Algorithm 2. Evaluate procedure

Input:
P: population
h: number of hidden nodes
it: number of training iterations
η: learning rate
m: momentum value
Output: P: Population evaluated

```
1  Begin
2    Foreach Chromosome ch ∈ P do
3      I=extractGeneIndexes(ch)
4      TestSet=GenerateTestSet(I)
5      TrainSet=GenerateTrainSet(I)
6      N=new Network(I, h, 1) // building an ANN with the I selected genes
7      N.train(Trainset, it, η, m) //Training N for it epochs with Trainset
8      Eval(N, TestSet, TMSE , ICI) //Evaluating ANN generalization ability
9      ch.fitness=(TMSE + ICI)/2 // computing ch fitness value
10   Return(P)
11 End
```

[1] A closed frequent itemset is a the largest set of items sharing the same transactions (objects).

[2] A minimal generator is a the smallest set of items sharing the same transactions (objects).

3 Experiments

3.1 Feature Selection Settings

Modeling tools based on a ANN can not be trained or assessed by a raw dataset. In our case, fortunately all the variable values have a numerical representation, and thus data have to be normalized. ANN algorithms require data to range within the unit interval. The chosen method for data normalization is the *linear transform scaling* [19]:

$$\nu_n = \frac{\nu_i - min\{\nu_1..\nu_n\}}{max\{\nu_1..\nu_n\} - min\{\nu_1..\nu_n\}} \tag{2}$$

Where ν_n and ν_i respectively represent the normalized and the actual values. This expression takes values and maps them them into corresponding values in the unit interval $[0, 1]$. The main advantage of the linear scaling is that it introduces no distortion to the variable distributions.

During carried out experiments, we have tested different values for each parameter. Table 1 summarizes neural network and genetic algorithm parameters, that permitted to obtain the best results.

Table 1. GA and ANN parameter settings

GA parameters		ANN parameters	
Parameter	Value	Parameter	Value
Number of generations	200	Number of iterations	250
Crossover probability(p_{cross})	80%	learning rate (η)	0.25
Mutation probability (p_{mut})	20%	Number of hidden nodes	10
Initial population size	30	Weights initialization range	$[-0.1..0.1]$
population size	20	Architecture	Feed forward fully-connected

Table 2. The 45 selected genes

Code	Probe set ID	average level	Code	Probe set ID	average level	Code	Probe set ID	average level
1	31469-s-at	43.54	2	32004-s-at	390.79	3	33647-s-at	432.65
4	34589-f-at	190.60	5	34600-s-at	411.39	6	36399-at	65.98
7	36411-s-at	538.70	8	32352-at	987.98	9	33495-at	70.75
10	33981-at	15.60	11	34037-at	13.90	12	34495-r-at	614.79
13	36770-at	11.76	14	37159-at	54.96	15	37483-at	146.19
16	39672-at	881.20	17	31853-at	111.69	18	31891-at	213.32
19	32672-at	42.20	20	33237-at	242.10	21	33334-at	110.69
22	34189-at	129.69	23	34664-at	91.65	24	36044-at	220.39
25	36927-at	17.32	26	39783-at	354.10	27	40449-at	22.81
29	40451-at	170.80	30	40485-at	201.10	31	40870-g-at	254.10
32	33344-at	40.20	33	34825-at	200.89	34	35775-at	14.90
35	37383-f-at	16806.40	36	39118-at	983.20	37	39494-at	432.87
38	39848-at	114.40	39	39922-at	57.70	40	40532-at	84.50
41	40958-at	132.45	42	32583-at	951.20	43	33144-at	53.87
44	942-at	65.21	45	323-at	98.43			

Starting with the previously defined parameters, we obtained a highly compact set genes whose size is by far lower than the initial number of genes, *i.e.*, more than twelve thousands. Table 2 sketches the 45 genes retained from among more than twelve thousands. Even though, more compact gene sets were obtained, the retained gene set achieves high generalization performance on test set: around 93% of accuracy.

3.2 Generic Association Rule Extraction

Table 3 illustrates cardinalities of the different generic basis extracted from the discretized "54×12 566" matrix for an absolute *minsup* value equal to 1 patient. Indeed, extraction context matrix has been translated into a boolean context by considering that a gene is over-expressed in a patient whenever his expression value level is greater than or equal to its expression level average at the different patients for the same gene.

Table 3. Extraction of generic association rules from the initial context

minconf	$\mathcal{IGB}$	($\mathcal{GBE}$, $\mathcal{TGBA}$)	$\frac{\mathcal{IGB}}{(\mathcal{GBE},\mathcal{TGBA})}$
0.05	1058829	6511866	0.162
0.3	5887121	6420922	0.916
0.5	5920969	6381928	0.927
0.7	6348233	6374305	0.995
1	999848	999848	999848

Table 3 shows important profits in terms of compactness of the $\mathcal{IGB}$ basis. Indeed, the third column of Table 3 shows that the ratio between the cardinality of $\mathcal{IGB}$ and that of ($\mathcal{GBE}$, $\mathcal{TGBA}$) ranges between 0.162 and 1.

Table 3 points out that the unmanageably highly sized association rule sets makes the perusal of knowledge ineffective. To palliate such drawback, we applied a feature selection process we retrieved only 45 "interesting" genes. From the selected genes, we constructed a binary $\mathcal{K}'$ context composed of 47 columns (45 genes, complete remission and relapse) and 54 rows (patients). Table 4 illustrates the cardinalities of the different generic basis. From Table 3 and Table 4, we can

Table 4. Extraction of generic association rules from filred context

minconf	$\mathcal{IGB}$	($\mathcal{GBE}$, $\mathcal{TGBA}$)	$\mathcal{NRR}$	$\frac{\mathcal{IGB}}{(\mathcal{GBE},\mathcal{TGBA})}$	$\frac{\mathcal{IGB}}{\mathcal{NRR}}$
0.05	852	3683	1974	0.231	0.431
0.3	3063	3432	1803	0.892	1.698
0.5	2398	2928	1422	0.818	1.686
0.7	1187	1336	605	0.888	1.961
1	850	850	24	1	35.416

conclude that the number of the generic association rules considerably decreased and this may permit to ameliorate the quality of the knowledge exploitation.

From the extracted generic rules of $\mathcal{IGB}$, we selected those whose conclusion part at least contains *complete remission / relapse*. Indeed thanks to the "Augmentation" axiom defined in [14], it is possible to straightforwardly derive "classification rules", i.e., rules whose the conclusion part refers to the class attribute. For example, from the post- feature selection process extracted generic rules– whose a sample is sketched by Figure 3– one may remark the following rule "22/129.69, 33/200.89 ⇒ 26/354.10, `Complete Remission`". From such rule, we can derive the following classification rule: "22/129.69, 33/200.89,26/354.10 ⇒ `Complete Remission`". This may permit to easily identify prognosis-associated genes. In order to facilitate the interpretation of the generic rules, we colored the patient situation (the green color corresponds to the complete remission, whereas the red one corresponds to the relapse). Also, it is important to mention that under an explicit request from experts, we decorated genes within the extracted rules by statistical information. This information represents the average of minimal expression level for each gene. Such information was considered of paramount importance by biologists since they were interested in checking the presence or absence of a given gene in conjunction with a significant signature appearance level.

22/129.69	26/354.10	===>	Complete Remission	Support : 6	Confiance : 1.000000			
22/129.69	33/200.89	===>	26/354.10	Complete Remission	Support : 5	Confiance : 1.000000		
2/390.79	7/538.70	===>	27/624.79	Complete Remission	Support : 5	Confiance : 1.000000		
2/390.79	26/354.10	===>	Complete Remission	Support : 5	Confiance : 1.000000			
28/170.80	35/16806.4	40/84.50	===>	Relapse	Support : 4	Confiance : 1.000000		
28/170.80	33/200.89	40/84.50	===>	Relapse	Support : 4	Confiance : 1.000000		
26/354.10	27/624.79	30/201.10	===>	Complete Remission	Support : 4	Confiance : 1.000000		
24/220.39	33/200.89	40/84.50	===>	Relapse	Support : 4	Confiance : 1.000000		
19/42.20	24/220.39	33/200.89	===>	26/354.10	Relapse	Support : 4	Confiance : 1.000000	
22/129.69	30/201.10	===>	4/190.60	26/354.10	33/200.89	Complete Remission	Support : 4	Confiance : 1.000000
4/190.60	30/201.10	33/200.89	===>	22/129.69	26/354.10	Complete Remission	Support : 4	Confiance : 1.000000
4/190.60	26/354.10	30/201.10	===>	22/129.69	33/200.89	Complete Remission	Support : 4	Confiance : 1.000000
4/190.60	22/129.69	26/354.10	===>	30/201.10	33/200.89	Complete Remission	Support : 4	Confiance : 1.000000
4/190.60	22/129.69	33/200.89	===>	26/354.10	30/201.10	Complete Remission	Support : 4	Confiance : 1.000000
19/42.20	22/129.69	===>	26/354.10	Complete Remission	Support : 4	Confiance : 1.000000		
19/42.20	21/110.69	28/170.80	===>	26/354.10	Relapse	Support : 4	Confiance : 1.000000	
17/111.69	21/110.69	31/254.10	===>	Relapse	Support : 4	Confiance : 1.000000		
16/881.20	24/220.39	===>	40/84.50	Relapse	Support : 4	Confiance : 1.000000		
16/881.20	40/84.50	===>	24/220.39	Relapse	Support : 4	Confiance : 1.000000		
7/538.70	28/170.80	33/200.89	===>	Relapse	Support : 4	Confiance : 1.000000		
2/390.79	26/354.10	27/624.79	===>	7/538.70	Complete Remission	Support : 4	Confiance : 1.000000	
2/390.79	7/538.70	26/354.10	===>	27/624.79	Complete Remission	Support : 4	Confiance : 1.000000	
2/390.79	4/190.60	5/411.39	28/170.80	===>	Relapse	Support : 4	Confiance : 1.000000	
24/220.39	39/57.70	===>	33/200.89	Relapse	Support : 3	Confiance : 1.000000		
28/170.80	38/114.40	===>	27/624.79	Relapse	Support : 3	Confiance : 1.000000		
26/354.10	34/14.90	===>	28/170.80	Relapse	Support : 3	Confiance : 1.000000		
31/254.10	42/951.20	===>	21/110.69	26/354.10	33/200.89	Complete Remission	Support : 3	Confiance : 1.000000
26/354.10	33/200.89	42/951.20	===>	21/110.69	31/254.10	Complete Remission	Support : 3	Confiance : 1.000000
21/110.69	42/951.20	===>	26/354.10	31/254.10	33/200.89	Complete Remission	Support : 3	Confiance : 1.000000
31/254.10	40/84.50	===>	17/111.69	Relapse	Support : 3	Confiance : 1.000000		
4/190.60	28/170.80	39/57.70	===>	Complete Remission	Support : 3	Confiance : 1.000000		

Fig. 3. The extracted rules

4 Conclusion

Under some number of hypothesis, generic association rules can constitute a gene annotation framework based on a strong correlation clustering. However, and even though they are compact, their high size can hamper their exploitation by experts.

In this paper, we proposed a novel approach towards filtering the most "predictive" compact set of genes. This approach, firstly uses genetic algorithms to filter out significant set of genes. Second, using this compact, we extracted reasonably sized generic association rules. Carried out experiments on a benchmark dataset showed the potential benefits of such approach. Indeed from more twelve thousands genes (possibly from which we may extract millions of generic rules and one imagine the number of all extractable association rules), we selected only 45 gene. From such reduced set of gene, it was possible to straightforwardly extract classification rules by means of associated derivation axioms.

References

1. Wang, J., M.J.Zaki, Toivonen, H., Shasha, D.: Data Mining in Bioinformatics. Advanced Information and Knowledge Processing. Springer (2005)
2. Chen, Y.: Bioinformatics Technologies. Advanced Information and Knowledge Processing. Springer (2005)
3. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97** (1997) 273–324
4. Hall, M.A., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Eengineering **15** (2003)
5. Cornujols, A., Miclet, L., Kodratoff, Y., Mitchell, T.: Apprentissage artificiel : concepts et algorithmes. Eyrolles (2002)
6. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison Wesley (1989)
7. Trabelsi, A., Esseghir, M.A.: New evolutionary bankruptcy forecasting model based on genetic algorithms and neural networks. 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) (2005) 241–245
8. Liu, H., Li, J., Wong, L.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics **13** (2002) 51–60
9. Shang, C., Shen, Q.: Aiding classification of gene expression data with feature selection: A comparative study. International Journal of Computational Intelligence Reasearch **1** (2005) 68–76
10. Esseghir, M.A., Yahia, S.B., Abdelhak, S.: Localizing compact set of genes involved in cancer diseases using an evolutionary conectionist approach. In: European Conferences on Machine Learning and European Conferences on Principles and Practice of Knowledge Discovery in Databases. ECML/PKDD Discovery Challenge. (2005)
11. A. Narayanan, A. Cheung, J.G.E.K.C.V.: Artificial neural networks for reducing the dimensionality of gene expression data. Bioinformatics Using Computational Intelligence Paradigms. Springer Verlag **176** (2005) 191–211
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge (1986)
13. Zaki, M.J.: Mining non-redundant association rules. Data Mining Knowledge Discovery **9** (2004) 223–248

14. Gasmi, G., BenYahia, S., Nguifo, E.M., Slimani, Y.: $\mathcal{IGB}$: A new informative generic base of association rules. In: Proceedings of the Intl. Ninth Pacific-Asia Conference on Knowledge Data Discovery (PAKDD'05), LNAI 3518, Hanoi, Vietnam, Springler-Verlag (2005) 81–90
15. Kryszkiewicz, M.: Representative association rules and minimum condition maximum consequence association rules. In: Proceedings of Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD), 1998, LNCS, volume 1510, Springer-Verlag, Nantes, France. (1998) 361–369
16. Zaki, M.: Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery (2004) 223–248
17. Zaki, M.J.: Generating non-redundant association rules. In: Proceedings of the 6th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA. (2000) 34–43
18. Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., Stumme, G.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Proceedings of the International Conference DOOD'2000, LNAI, volume 1861, Springer-Verlag, London, UK. (2000) 972–986
19. Pyle, D.: Data Preparation for Data Mining. (1999)