Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective

Menolascina F, Tommasi S and Paradiso A
Clinical and Experimental Oncology Laboratory
National Cancer Institute
Bari, 70126, Italy
filippo.menolascina@gmail.com

Cortellino M, Bevilacqua V and Mastronardi G
Dept. of Electronics and Electrical Engineering
Polytechnic of Bari
Bari, 70125, Italy
bevilacqua@poliba.it

***Abstract* – In this paper we present a comparative study among well established data mining algorithm (namely J48 and Naïve Bayes Tree) and novel machine learning paradigms like Ant Miner and Gene Expression Programming. The aim of this study was to discover significant rules discriminating ER+ and ER- cases of breast cancer. We compared both statistical accuracy and biological validity of the results using common statistical methods and Gene Ontology. Some worth noting characteristics of these systems have been observed and analysed even giving some possible interpretations of findings. With this study we tried to show how intelligent systems can be employed in the design of experimental pipeline in disease processes investigation and how deriving high-throughput results can be validated using new computational tools. Results returned by this approach seem to encourage new efforts in this field.**

***Index Terms* – Ant Miner, Breast Caner, Decision Trees, Gene Expression Programming, Rule Induction.**

Supplementary material*: http://oncologico.bari.it/*

I. Introduction

Chromosomal aberrations have been showed to be frequently involved in human cancer development [1]. Genomic DNA alteration, i.e. loss or amplification of specific genes, in fact, can markedly rise the probability of carcinogenesis in healthy patients. Gene dosage becomes, in this context, a particularly interesting variable to be monitored in order to rise the effectiveness of early diagnosis in human tumours. Different kinds of approaches have been proposed to study such disorders; Fluorescent In Situ Hybridization (FISH) and Representational Difference Analysis (RDA) and Comparative Genomic Hybridization (CGH) [2]. The last is a powerful technique although its usefulness is greatly limited by intrinsic technical limitations that prevent it to become a comprehensive screening tool. However, recent advancements in technologies have allowed researchers to conjugate the strength of CGH and microarray platforms in Array Comparative Genomic Hybridization (aCGH) [3][4]. Results of aCGH screening are in the form of microarray images (Fig. 1); spot intensities are evaluated as ratios of fluorescent tag concentration and corresponding values are associated to specific probe copy number. Bacterial Artificial Chromosome (BACs) have been commonly used as probes in order to observe copy number changes of regions of the genome that share the same relative copy number on average. The resulting set of values for each patient can be analysed as a profile of genomic segments, as reported in Fig 2. For analysis purposes raw values are transformed applying $\log_2$(ratio) transform; this step is meant to give a theoretically 0 median for regions where no alteration occurred. On the other hand segments with positive means represents duplicated regions in the test sample genome and segments with negative means characterize deleted regions of the DNA. It is important to note that although the biological entity (copy number) is intrinsically discrete, the signal under investigation is considered as being continuous; this inconsistency is due to the fact that quantification of copy number levels is based on fluorescence measurement that is of an analogue source. The obtained profiles constitute quasi raw data; this is the starting point for all the following analysis steps that will guide the researcher to the extraction of useful knowledge about the disease under investigation.

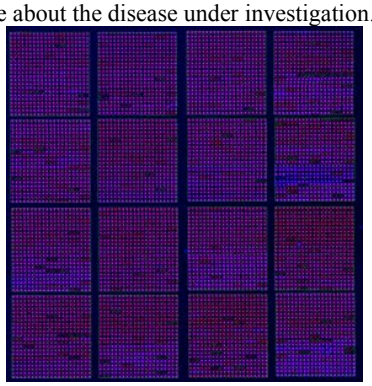

Fig. 1 Each spot in the array corresponds to a single BAC probe. Spot intensity associated to BAC clones is directly dependent on copy number levels of genes included in the clone, i.e. the more the spot is enriched with fluorescent tag, the higher the copy number level of the genes and the more severe the genomic alteration.

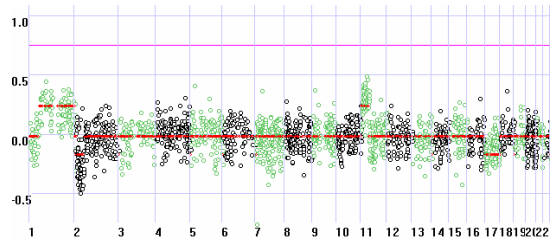

Fig. 2 Whole genomic profile of patient affected by BC. Regions with amplifications and deletions are clearly visible. It is even evident that this kind of approach can easily return a comprehensive snapshot of DNA copy number alterations in a single experiment.

Several diverse approaches to data mining in similar context have been proposed previously [5][6] however literature tends often to focus on the results as "abstract entities" only supported by statistical evidence. This approach has been demonstrated to be quite debatable due to the fact the statistical support doesn't always deterministically imply real coherence between results and biological processes. Moreover several questions in bioinformatics knowledge-extraction-flow management remain open; they mainly concern data reduction strategies, mining algorithms and results interpretation. The research of the last years put in evidence that no "global optimum" exists in the field of data analysis, and that, instead, several approaches seem to perform well in some restricted areas. On the other hand novel paradigms based on machine learning have been shown to return interesting results even when compared to well established statistical based methods. Here we present a new experimental pipeline that takes advantage of some of the most promising among emergent methods for rule induction (namely Ant Colony Optimization – Ant Miner - and Gene Expression Programming -GEP) establishing a comparative study with well known data mining algorithms (namely J48 and Naïve Bayes Tree). The choice of rule induction algorithm for bioinformatics and biomedical data mining will be explained in one of the next paragraphs. We further illustrate a feasible approach to data dimensionality compression based on statistical properties of features describing observations. We built a consensus criterion for predictive power estimation of each of the BACs sod the reduced input set using three main methods: *Student's T-test*, Receiver Operating Curve (*ROC*) and *entropy* (Kullback-Lieber divergence).

All the classifiers were built starting from this common point and trained on a fixed set of features. Results have been collected iterating multiple times non deterministic the algorithms like GEP and Ant Miner in order to overcome the intrinsic variability of similar algorithms and analysing top ranked rules. The validation of the results is double; on one hand we compared accuracies reached by systems and, contemporary we tried to show how biological perspective could be integrated in this kind of analysis using Gene Ontology [7].

The automated integration of background knowledge is fundamental to support the generation and validation of hypotheses about the function of gene products. One such source of prior knowledge is the Gene Ontology (GO), which is a structured, shared vocabulary that allows the annotation of gene products across different model organisms. The GO comprises three independent hierarchies: molecular function (MF), biological process (BP) and cellular component (CC).

Researchers can represent relationships between gene products and annotation terms in these hierarchies. Potentialities of GO in knowledge driven validation of the experimental results is an evident result of its design. In this work we propose a biological interpretation of the clusters of genes able to discriminate ER status in breast cancer subtype profiling.

Gene Ontology allowed to gain deeper insights in the biological mechanisms underlying the disease under investigation. Correlation of previously unconsidered genes with known BC biomarker emerged and pushed further investigation on these genes. With this study we tried to show how intelligent systems can be employed in the design of experimental pipeline in disease processes investigation and how deriving high-throughput results can be validated using new computational tools. Results returned by this approach seem to encourage new efforts in this field.

II. Materials and Methods

A. Samples

In this study we considered a cohort of 124 patients with breast cancer at different stages. Samples were collected and treated as reported in [8].

A statistical summary of the case set used in this study is presented in Table 1.

	Summary Statistics		
	All (n=124)	***ER Positive***	***ER Negative***
Age (*Kendall's tau b p* = 0.318)			
Young (<= 45 y.o.)	56 (45,5%)	33	23
Old (>= 70 y.o.)	66 (53,7%)	57	9
T Stage (*Kendall's tau b p* = 0.028)			
T1	31 (25,2%)	24	7
T2	59 (48,0%)	39	20
T3	8 (6,5%)	8	0
T4	20 (16,3%)	16	4
Grade (*Cramer's V p* = 0.325)			
1	15 (12,2%)	13	2
2	57 (46,3%)	45	12
3	35 (28,5%)	18	17
Missing	15 (12,2%)		
PgR Status (χ^2 *p* = 0.216)			
PgR Positive	58 (%)	37	21
PgR Negative	65 (%)	53	12

	All (n=124)	*ER Positive*	*ER Negative*
Proliferation (χ^2 $p = 0.196$)			
MIB Negative	18 (14,6%)	17	1
MIB Positive	105 (85,4%)	73	32

Tab. 1 Summary statistics of the dataset. In order to investigate the relationship between classes diverse metrics have been used. Cramer's V measures the strength of a relationship of two nominal variables when one or both have three or more levels or values; Kendall's Tau b is necessary when it comes to measuring strength of association if both variables are ordinal; χ^2 is used in contexts with two dichotomous or nominal variables. No evident relationship is noticeable between subclasses.

B. Algorithms

In this section we will give a brief description of the algorithms and the paradigms they are inspired to. In particular we will analyse ACO in the flavour of Ant Miner [9][10][11] and Gene Expression Programming. In addition we will examine two well established approaches to data mining: J48, an evolution of C4.5, and Naïve Bayes Tree. There's an intrinsic dichotomy in classification problems in medicine that concerns the main objective of the research. It could be argued that the only goal of the study is to develop a system that is able to impute correctly cases to classes, in this case we assume a "black-box" abstraction of the system being developed (Artificial Neural Networks or Support Vector Machines, for example). Similar kind of algorithms take some inputs and return some outputs; they can reach a variable level of accuracy but they will not enrich the human knowledge of the process under investigation.

This is a key point in the biomedical context: physicians often want to understand the way the classifier is behaving in order to judge its performances. This is a quite interesting perspective: underlying their interest there's the willing of gaining a deeper knowledge of the biological process for trying to interpret the results returned by the system. This is a peculiar aspect of the biomedical field in which a percent point in the classifier accuracy can decide the survival of a human being. Another model is then needed in order to address these requests. The second approach, then, gives a deeper insight into the problem adding to the prediction a clear scheme it followed in order to reach the prediction. These schemes are called *rules* and the process of rule extraction from a dataset is called *rule induction*.

Several different rule induction approaches have been proposed however one of the most representative field of research in this area is the one of trees. Trees are common structures in informatics and evidently they can be successfully used in rule representation. Nodes of the tree can in fact assumed to be features of the dataset and branches can be interpreted as partition of the dataset that satisfy a given discriminating condition, as represented in Fig. 3. One of the most famous algorithm in the field of data mining that builds trees is C4.5; originally developed by Quinlan [9] this is one of the standard algorithms for translating raw data in useful knowledge. Rule induction systems are currently employed in several different environments ranging from loan request evaluation to fraud detection and medicine.

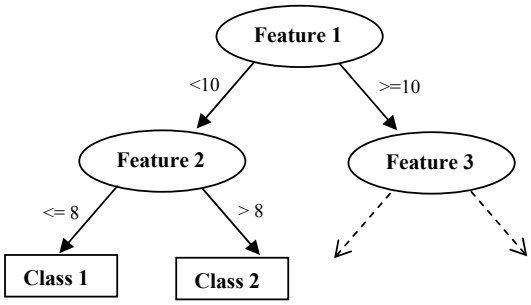

Fig. 3 Example of a classification tree. Two rules can be inferred from this structure being: 1) IF Feature 1 < 10 AND Feature 2 <= 8 THEN Class 1; 2) IF Feature 1 < 10 AND Feature 2 > 8 THEN Class 2.

It is then evident the advantage of rule mining systems over black-box systems when it comes to classification interpretation; in the biomedical context the discovery of the rules can ease the complexity for translating the complex raw data into relevant and clinically useful diagnostic or prognostic knowledge. However several different approaches to rule mining have been proposed in the recent years, here we present a comparative study of two of these paradigms based on computational intelligence models; they are the Ant Miner algorithm [10], inspired by the collective intelligent behaviour of ants in foraging tasks, and Gene Expression Programming [11], a hybrid model that mixes genetic algorithms and genetic programming in order to grow algebraic expressions which use features as variables. We compared the results of this systems with J48, an evolution of the C4.5, an entropy based data mining algorithm, and Naïve Bayes Tree, a tree construction algorithm that uses the paradigm of Bayes theory with strong independence assumption. An outlook of all of these algorithm is presented in the following sections of the paper.

a. J48 Classifier

J48 classifier forms rules from pruned partial decision trees built using C4.5's heuristics. C4.5 is Quinlan's most recent non- commercial tree-building algorithm. The main goal of this scheme is to minimize the number of tree levels and tree nodes, thereby maximizing data generalization. It uses a measure taken from information theory to help with the attribute selection process. For any choice point in the tree, it selects the attribute that splits the data so as to show the largest mount of gain in information. The J48 classifier described above builds a C4.5 decision tree. Each run of J48 it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier, are all part of that instantiation of the J48 class. J48 class does not actually contain any code for building a decision tree. It includes references to instances of other classes that do most of the work. It also combines the divide-and-conquer strategy for

decision tree and separate divide-and-conquer one for rule learning. Such approach adds flexibility and speed.

b. Naïve Bayes Tree

Naïve Bayes Tree is a hybrid between decision trees and Naïve Bayes. This algorithm creates trees whose leaves are Naïve Bayes classifiers for instances that reach the leaf. When constructing the tree, cross-validation is used to decide whether the node should be split further or a Naïve Bayes model should be used instead. The algorithm is similar to the classical recursive partitioning schemes, except that the leaf nodes created are Naïve Bayes categorizers instead of nodes predicting a single class. A threshold for continuous attributes is chosen using the standard entropy minimization technique, as is done for decision-trees. The utility of a node is computed by discretising the data and computing the 5-fold cross-validation accuracy estimate of using Naïve-Bayes at the node. The utility of a split is the weighted sum of the utility of the nodes, where the weight given to a node is proportional to the number of instances that go down to that node. Intuitively the algorithm tries to approximate whether the generalization accuracy for Naïve-Bayes classifier at each leaf is higher than the single Naïve-Bayes classifier at the current node. To avoid splits with little value, we define a split to be significant if the relative (not absolute) reduction in error is greater than 5% and there are at least 30 instances in the node.

c. Ant Miner

Ant based algorithms or ant colony optimization (ACO) have been applied successfully to combinatorial optimization problems. More recently Parpinelli and colleagues have applied ACO to data mining classification problems, where they introduced a classification algorithm called Ant Miner. The goal of Ant miner is to extract classification rules from data [REF: Parpinelli 2002, 12] this is accomplished by leaving agents (ant) exploring the space of attributes looking for best combination of antecedents that predict a given class. An overview of the Ant Miner algorithm is given in Fig. 4.

```
TS = all training cases;
WHILE (No. of cases in TS > max_uncovered_cases)
        i=0;
        REPEAT
        i=i+1; Ant_i incrementally constructs a rule;
                Prune the just constructed rule;
                Update the pheromone of the trail
                followed by Ant I;
        UNTIL (i ≥ No_of_Ants) or (Ant_i constructed the same rule as the previous No_Rules_Converg-1 Ants)
        Select the best rule among all constructed rules;
        Remove the cases correctly covered by the selected rule from the training set;
END
```

Fig. 4 Pseudocode of the Ant Miner algorithm in Parpinelli's implementation.

Ant Colony Algorithms have been recently used in classification problems in bioinformatics by Chan and Freitas in [13].

d. Gene Expression Programming for Rule Mining

Gene expression programming was first proposed by C. Ferreira in [25] as an alternative of Genetic Programming. GEP uses linear chromosomes of fixed length which are afterwards expressed in non linear entities of different size and shapes (expression trees). The genotype/phenotype translation is performed with a depth first visit of the tree, as example consider the following expression tree in Fig. 5:

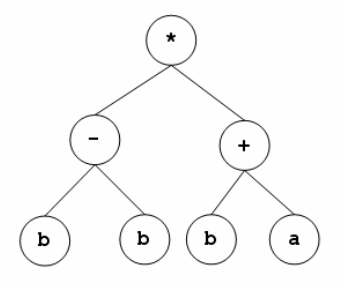

Fig. 5 Example of expression tree.

The linear representation is the list of nodes from top to bottom, left to right: *-+bbba. The opposite process is also simple: the first symbol is the root, successive symbols are attached in sequence below every function node regards to his arity . The assemblage stops when no more symbols are needed. Genes are structured in two parts: head, in witch terminal and non terminal symbols could be present, and tail in witch only non terminal symbols are allowed. In order to ensure the validity of any expression, the tail length must be at least:

(1) $t = (n - 1) * h + 1$ where n is the maximum number of arguments of a function and h is the head length. It can happen that not all symbols of a gene are expressed in the relative tree, for example consider a gene with h = 3, n = 2 and then, according with (1) t = 4:

```
1234567
/a+baca
```

Notice that symbols 6 and 7 are not present in the tree because the expression process stops before they are parsed. The expressed part of chromosome is named ORF (open read frame). The non coding region is also important because it can become part of the ORF during the evolution. A chromosome can also contains more than one gene, if this is the case genes can be linked with a simple predefined operator like in [25] or with a generic function coded directly in the gene after the tail. GEP uses common genetic operators (selection, mutation and crossover) plus other more specific operators: transposition and gene recombination. We focus the attention on the evolution of classification rules with GEP [26]. Given a set of examples each of witch described by a vector of n numerical features and a class of membership, the goal is to find for every class C_i one or more functions $f_i(x)$, where x is the features vector, such that if $f_i(x) > 0$ then x belongs to class C_i . This approach was used by Nelson et al. in [26]

for discovering compact classification rules. The learning is supervised, and, for each class C_i, a set of rules is evolved: the first rule is trained over all training set, then all examples of class C_i covered by this rule are eliminated and a new rule is evolved over the remaining examples. The process stops when all positive examples (i.e. belonging to C_i) are covered.

The fitness used for individual selection is:

$$fitness(R) = \begin{cases} 0 & if \quad consig(R) < 0 \\ consig(R) * \exp(compl(R) - 1) & otherwise \end{cases}$$

Where:

$$consig(R) = \left(\frac{p}{p+n} - \frac{P}{P+N} \right) * \frac{P+N}{N} \quad compl(R) = \frac{p}{P}$$

P and N are respectively the total number of positive and negative examples in the training set, p and n are respectively the amount of true positives and false positives. If the rule is perfect, (p = P and n = 0) the fitness returns one, in the case of a random choice (p = P/2 and n = N/2) the fitness returns zero. If training set is noisy, complex rules that fit a small number of examples maybe appear. To avoid overfitting a stop criterion based on Minimum Description Length principle is used, in particular for each rule evolved, the length of rule set L(H) is updated as follow:

$$L(H) = L_{exception}(H) + w * L_{theory}(H)$$

Where:

$$L_{exception}(H) = \log_2 \binom{n}{N_{fp}} + \log_2 \binom{N-n}{N_{fn}}$$
$$L_{theory} = L_{theory} + \log_2 (N_c) * L(R_i)$$

After a rule is evolved, if the total length L(H) is greater than the previous one, the rule is discarded and evolution stops, otherwise, it is added to the rule set and, if there are still positive examples, a new rule is evolved.

When the complete rule set is evolved, a new pruning phase is performed: rules are ordered by fitness and the best is added to final ordered list of rules, than all correct classified examples covered by this rule are removed from the training set. Fitness is recalculated for remaining rules and the process was repeated until no more examples remains. Finally a default class is assigned to those class that has the greatest number of unclassified example. This procedure attempts to avoid conflicts between rules.

III. Gene Selection Criteria

The feature selection stage is one of the most delicate in the whole micro-array experimental pipeline. In a common array based experiment it's not uncommon to handle a gene set of the order of thousands. Of course it is obvious that such a feature set can be greatly optimized, eliminating redundancy of co-regulated genes, for example, or considering subsets of genes that minimize inter correlation. Many different approaches are documented in literature; the most recent contribute to this field of optimal feature set finding comes from [14]. Many relevant suggestions can be found in this work in particular about the covariance structure of data and its impact on the optimality of feature set. Other feasible approaches include sensitivity analysis by removing attributes, proportion correct use in rules, ratio of features Between-category to Within-category sums of squares, Signal-to-Noise scores in Onve-versus-Rest or One-versus-All fashion, Kruskall-Wallis non parametric test (ANOVA) and number of appearances in models [16] [18] [19]. However the scientific community seems to agree that the "optimal feature set" simply doesn't exist but, instead, it should be measured on the single classification approach and, in general, on the single experiment [15]. For this reason we developed a consensus scheme for attribute selection that takes advantage of three well established statistical methods, they are: Student's T-Test, Receiver Operating Characteristic and Entropy (Kullback-Lieber divergence). T-Test checks for mean of a distribution and allow to establish a comparison of diversity between two populations through mean comparison; this test returns a value that can be easily translated in the probability that the sets of data are drawn from the same distribution or from different distribution with the same means. IN ROC the Area Under The Curve is estimated as an indicator of class separation; the more separated the classes, the higher the AUC. Kullback-Lieber divergence, instead, is a principle drawn for information theory that accounts for inner information contained in each attribute being a good metrics for more expressive attributes selection. All of these techniques can be used to compile a ranking of the features that accounts for the power of a single attribute to discriminate between the output classes. All of the 2464 BAC values for each of the 124 cases where processed the outcome being ER status; using these algorithms three rankings have been obtained. A new global ranking has been compiled using the three positions of each clone as an indicator of its discriminating power. This strategy has been employed in order to overcome the limitations of the single methods and to gain a deeper insight into the data structure and information distribution. In addition, as reported in [17], it should be considered that using a single viewpoint for relevance estimation can results in unbearable bias in results. Bonferroni adjustment ahs been employed in order to correct the statistics for multiple comparison. The former first forty clones were selected for the following analysis stages.

IV. Experiments

In this section we present the experiments performed on the four machine learning algorithms for aCGH based Breast Cancer Subtypes profiling. The subtype chosen for this research was ER status, being two the possible outcomes: ER+ or ER-. ER status is one of the parameters

mostly used as breast cancer characterizer because of its high correlation with aggressiveness of the pathology.

First we will show dataset descriptions and the preprocessing method. Next we will present the results of the experiments carried out.

A.. Data and Preprocessing

The output of aCGH array scanning has been converted in $\log_2(R_1/R_2)$ where R_1/R_2 indicates the ratio of the two fluorescent tags; this is a common pre-processing of the data that tries to overcome the bias introduces by the fact that lost and normal BACs are theoretically compressed in the interval [0,1], and, on the other hand, amplifications can vary in the range]1, ∞[. The matrix obtained is then composed by 119 cases each of which is defined by 2464 log ratios.

At this point some missing values exist in the dataset (also indicated as *NaN*, Not a Number)and a decision about these and the BACs they belong is needed.

Some approaches to missing values handling tend to simply eliminate those features that contain missing values; this, obviously, brings inevitably to some loss of information. Another approach consists in imputing missing values using other information; the most simple method imputes a missing value using the mean (or median) of the distribution of the single BAC to all the missing values; it is evident that if a single case out of all contains a value lost for all of the others, this methods will impute this single value to all of the cases leading to a strong bias in data. If the cases are two and each of the two belongs to one of the classes under investigation it's clear that mean imputation, in this case, will make powerful gene selection criteria like Wilcoxon test or Student's T-test absolutely inadequate. For these reasons we choose a hybrid approach to missing value imputation: we firstly removed all the BACs that were present in the 33% of the cases. Then we used Collateral Missing Value Estimation (CMVE) as described in [20].

As final step we applied gene entropy filter [21] to the dataset obtaining a matrix 119 by 2218. These set of genes has been used as input for the gene selection algorithm.

B. Perfomance Measures

As a performance measure we used global accuracy of the systems and Kappa-Statistic.

Kappa-Statistic is commonly used as a measure of the advantage of the classifier under investigation over a random classifier.

Global accuracy is defined by the ratio of the number of correct predictions and the number of all predictions as explained in the Eq. 1:

$$Glob.Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (1)$$

where TP stands for True Positives, TN True Negatives, FP for False Positives, and FN for False Negatives.

Ant Miner and Gene Expression Programming tests were repeated 100 times to account for intrinsic variability of the results obtained. The rules and antecedents with higher frequencies were selected as significant till a p value of 0.05. The accuracy results are expressed in terms of medians of the values extracted in the case of Ant Miner and GEP. The strategy for training and validation selected was the K-Fold cross-validation with K=10.
The results are shown in Fig. 6.

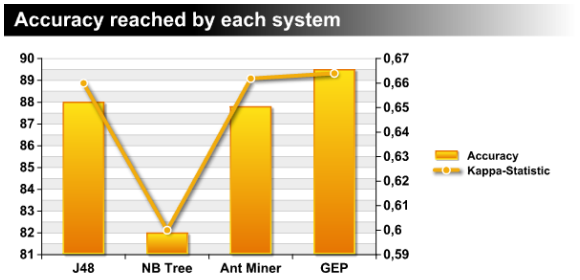

Fig. 6 Accuracy and Kappa-Statistic for each system.

Results returned by the experiments carried out show a quite clear situation; the performances of J48, Ant Miner and GEP algorithms are comparable, the last having a small advantage over the others. The global level of accuracy reached by these systems nears the 90%; this means that 9 cases over 10 are correctly covered by the rules this systems have generated. This outlook on the systems' capabilities is based on statistical validation and no further interpretation of coherence between the results and actual biological processes can be inferred. However confidence with results grows strongly with the understanding of the mechanism underlying decisions. For these reasons and the nature of the research we tried to validate the results using a knowledge driven approach. We used Gene Ontology (GO) in order to discover interesting patterns in the rules extracted by the systems.

V. Biological Validation

In order to understand the meaning underlying the rules extracted by the systems and if a logical coherence exists between the results an the expected process involved GO has been used in this validation stage.

Only the top performing rules have been analysed; J48, Ant Miner and Gene Expression Programming algorithms results, then, have been considered for further analyses. In Tab 2 most significant rules returned by each system are reported.

In order to adequately supply GO Miner [22] with a set of genes it is able to handle we translated the probe set of BACs of the array in corresponding genes' HUGO official names using Matchminer [23]. After having obtained the complete set of translations the GO analysis could be carried out.

Statistical assessment of significance of GO Terms has been accomplished using False Discovery Rate (*FDR*, [24] threshold $q < 0.2$) with a number of permutations equal to 100.

	Rules
J48	IF CTD-2079J2 <= -0.13249 THEN ER- IF CTD-2079J2 > -0.13249 AND RP11-542B5 <= 0.06908 AND RP11-59D4 <= -0.021134 THEN ER+ IF CTD-2079J2 > -0.13249 AND RP11-542B5 <= 0.06908 AND RP11-59D4 > -0.021134 AND RP11-53F9 <= 0.012435 AND RP11-69A18 <= -0.001621 AND RP11-6L19 <= 0.052644 THEN ER- IF CTD-2079J2 > -0.13249 AND RP11-542B5 <= 0.06908 AND RP11-59D4 > -0.021134 AND RP11-53F9 <= 0.012435 AND RP11-69A18 <= -0.001621 AND RP11-6L19 > 0.052644 THEN ER+ IF CTD-2079J2 > -0.13249 AND RP11-542B5 <= 0.06908 AND RP11-59D4 > -0.021134 AND RP11-53F9 <= 0.012435 AND RP11-69A18 > - 0.001621 THEN ER+ IF CTD-2079J2 > -0.13249 AND RP11-542B5 <= 0.06908 AND RP11-59D4 > -0.021134 AND RP11-53F9 > 0.012435 THEN ER+ IF CTD-2079J2 > -0.13249 AND RP11-542B5 > 0.06908 AND GS-561N1 <= 0.059564 THEN ER- IF CTD-2079J2 > -0.13249 AND RP11-542B5 > 0.06908 AND GS-561N1 > 0.059564 THEN ER+
Ant Miner	IF RP11-45L17 = lost THEN ER+ IF RP11-77O20 = ampl AND RP11-172K14 = lost THEN ER- IF RP11-116D2 = ampl AND RP11-327F6 = ampl THEN ER- IF RP11-45L17 = norm THEN ER+ Default: ER+
GEP	if ((((RP11-180G13 - RP11-162I18) - RP11-110L8) + (RP11-162I18 > RP11-78A18)) > RP11-5B23) then CLASS ER+ if IF(IF(RP11-174I22, RP11-59D4, RP11-2I4),(RP11-13O21 ! RP11-182E4),(RP11-53F9 > RP11-233E5)) then CLASS ER+ DEFAULT CLASS ER-

Tab. 2 Most significant rules extracted by top performing systems.

GO analysis has been carried out on the result of each of the algorithms in order to firstly assess intrinsic coherence with expected findings and then to observe if some kind of extrinsic consensus could be established among all the results. For the J48 algorithm it should be noticed that the best discriminating clone, namely CTD-2079J2, contains an onco-suppressor gene currently being under investigation. Moreover some interesting pathways have been discovered like 'cell-to-cell signaling'. Analysing the genes obtained from the translation of the BACs highlighted by Ant Miner algorithm we observed that even in this case cell regulation and signaling pathways resulted enriched with statistical significance. An interesting pathway discovered in this analysis is the one activated by APOB gene involved in 'steroid metabolism'. It should be noted that both these analyses individuated C10orf68 as a good discriminator. Deeper researches are being carried out on these sequence in order to establish candidate roles in the estrogen related pathways. The interpretation of rules extracted by GEP is not straight as J48's or Ant Miner's is; in this cases the use of GO can ease the interpretability of rule. In our case we observed 'cell differentiation', 'development' and 'cell-to-cell signaling' GO Term enriched by genes; in particular ID2 gene is believed to inhibit the functions of basic helix-loop-helix transcription factors in a dominant-negative manner by suppressing their heterodimerization partners through the HLH domains. This protein may play a role in negatively regulating cell differentiation. Finally we observed a strong correlation of GO Terms discovered by the three top performing algorithms. There's an evident coherence among the results returned meaning that all the three approaches have discovered overlapping when not similar properties of the dataset.

Further researches are being carried by the Clinical and Experimental Oncology Laboratory of NCI in order to uncover hidden properties of the results presented here. In particular FER role in ER status determination and PDGF/VEGFC interactions in breast cancer and related metastatic processes development.

VI. Conclusions

In this work we proposed a comparative study of novel machine learning paradigms trying to evaluate results both from the statistical and the biological perspectives. Data mining techniques can greatly help experts in extracting useful knowledge from databases where huge amounts of data are stored. This task becomes particularly delicate in the biomedical field where, usually, this already problematic situation is worsened by the high disproportion between the number of features and records.For these reasons we tried to estimate how different models of classifiers performed. We focused our research on systems generating decision trees or rules because of specific requests experts made in terms of system behavior interpretability and reliability estimation. All the systems showed good performances, however J48, Ant Miner and Gene Expression Programming algorithm were characterized by comparable and higher accuracy levels when compared with Naïve Bayes Tree approach. Multiple runs of the non deterministic algorithms were carried out in order to overcome the variable nature of the results returned by these approaches. Finally the global level of accuracy and Kappa-Statistic calculated over these three systems allowed to be moderately confident with

rules generated and their coverage. Biological interpretation of the results, carried out using GO, showed that all the top performing methods extracted BACs containing genes belonging to same or biologically correlated GO Terms. Moreover several interesting pathways and genes have been highlighted whose function and role in breast cancer ER status determination is currently under investigation. Some of the GO Terms extracted in this analysis resulted even directly involved in estrogen metabolism like it should be expected; the contemporary presence of estrogen metabolism related genes and PDGF/VEGF family of endothelial growth factors (known to be involved in angiogenesis and vascularisation of tumours) seem to be coherent with known correlation between ER status and tumor aggressiveness. We can conclude that the experimental pipeline described appears to return results reasonably correlated with processes expected to result highlighted. Novel biologically inspired data mining techniques, then, seem to be competitive complementary tools in cancer research being GEP, probably, the less explored. This approach has a strong expressive power that, unfortunately, is noticeably limited by the way rules are currently represented. More efforts should be made in this way in order to simplify rule interpretation in GEP, then allowing all the potential of this technique to be expressed. Further studies are being carried out in order to optimize the number of features to include in the training set and on the algorithms to be used according to the suggestions collected in [14]. Other studies currently under investigation include sensitivity analysis on the input parameters of both Ant Miner and GEP classifiers and the use of fuzzy rules to model biological mechanisms underlying a complex process like breast cancer is.

REFERENCES

[1] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO., Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12963-8.

[2] Beheshti B, Park P, Braude I, Squire J: Molecular Cytogenetics: Protocols and Applications Humana Press; 2002.

[3] Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: Matrix-based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. Genes, Chromosomes and Cancer 1997, 20:399-407.

[4] Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C,Kuo W, Chen C, Zhai Y, Dairkee S, Ljung B, Gray J: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 1998, 20:207-211.

[5] Sidhu, A. S., Kennedy, P. J., Simoff, S., Dillon, T. S. & Chang, E. (2007) Knowledge Discovery in Biomedical Data facilitated by Domain Ontologies. IN ZHU, X. & DAVIDSON, I. (Eds.) Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data. Hershey, Idea Group Inc. – BKC

[6] Kurra G, Niu W, Bhatnagar R, Mining microarray expression data for classifier gene-cores, BIOKDD 2001

[7] http://www.geneontology.org

[8] S. Tommasi et al. in preparation

[9] Quinlan, J.R. C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA. 1993.

[10] Parpinelli, R.S., Lopes, H.S., Freitas, A.A. "Data mining with an ant colony optimization algorithm". IEEE Transactions on Evolutionary Computation, special issue on Ant Colony Algorithms, v. 6, n. 4, p. 321-332, August, 2002.

[11] Ferreira, C., Gene Expression Programming in Problem Solving. In R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, and F. Hoffmann, eds., Soft Computing and Industry: Recent Applications, pages 635-654, Springer-Verlag, 2002

[12] R. Parpinelli, H. Lopes and A. Freitas, An Ant Colony Algorithm for Classification Rule Discovery, Data Mining: a Heuristic Approach, pp. 191-208. London: Idea Group Publishing, 2002.

[13] A. Chan and A. A. Freitas, A New Ant Colony Algorithm for Multi-Label Classification with Applications in Bioinformatics, GECCO 2006.

[14] E. R. Dougherty and M. Brun, On the Number of Close-to-Optimal Feature Sets, Cancer Informatics 2006: 2 189-196

[15] L. Ein-Dor et al "Outcome segnature genes in breast cancer: is there a inique set?" Bioinformatics 21: 171-178

[16] Gopalakrishnan et al, Rule Learning for Disease-specific Biomarker Discovery from Clinical Proteomic Mass Spectra, KDLL 2006.

[17] Li et al, Discovery of significant rules for classifying cancer diagnosis data, Bioinformatics 2003 2: 93-102.

[18] A. Stanikov et al, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics Advance Access, September 16 (2004).

[19] T. Golub et al, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, vol. 286, October 15 (1999).

[20] M. S. B. Sehgal, I. Gondal, and L. Dooley, "Collateral Missing Value Imputation: a new robust missing value estimation algorithm for microarray data," Bioinformatics, vol. 21(10), pp. 2417-2423, 2005

[21] Kohane I.S., Kho A.T., Butte A.J. (2003), Microarrays for an Integrative Genomics, Cambridge, MA:MIT Press.

[22] Barry R. Zeeberg, Weimin Feng, Geoffrey Wang, May D. Wang, Anthony T. Fojo, Margot Sunshine, Sudarshan Narasimhan, David W. Kane, William C. Reinhold, Samir Lababidi, Kimberly J. Bussey, Joseph Riss, J. Carl Barrett, and John N. Weinstein, A Resource for Biological Interpretation of Genomic and Proteomic Data. Genome Biology, 2003 4(4):R28

[23] Kimberly J Bussey, David Kane, Margot Sunshine, Sudar Narasimhan, Satoshi Nishizuka, William C Reinhold, Barry Zeeberg, Ajay and John N Weinstein, MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biology, 2003 4(4):R27

[24] Benjamini, Y., and Hochberg Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.C.Ferreira. Gene expression programming: a new adaptive algorithm for solving problems. Complex system, 13(2):87_129, 2001.

[25] Chi Zhou, Weimin Xiao, Peter C. Nelson, and Thomas M. Tirpak, 2003. Evolving Accurate and Compact Classification Rules with Gene Expression Programming. IEEE Transactions on Evolutionary Computation, Vol. 7, No. 6, pages 519-531.

[26] C.Ferreira. Gene expression programming: a new adaptive algorithm for solving problems. Complex system, 13(2):87_129, 2001.