International Journal of Hybrid Intelligent Systems 5 (2008) 1–5
IOS Press

Fuzzy rule induction and artificial immune systems in female breast cancer familiarity profiling

Filippo Menolascina[a,*], Roberto T. Alves[c], Stefania Tommasi[b], Patrizia Chiarappa[b], Myriam Delgado[c], Vitoantonio Bevilacqua[a], Giuseppe Mastronardi[a], Alex A. Freitas[d] and Angelo Paradiso[b]
[a]*Department of Electronics and Electrical Engineering, Polytechnic of Bari, 2 Via E. Orabona 4, 70125. Bari, Italy*
[b]*Clinical and Experimental Oncology Laboratory – NCI, 1 Via Hahnemann 10, 70126. Bari, Italy*
[c]*Federal Technological University of Paranà – UTFPR, 3 Av. 7 de setembro, 3165. Curitiba, Brazil*
[d]*Computing Laboratory – University of Kent, CT2 NF. Canterbury, UK*

Abstract. Genomic DNA copy number aberrations are frequent in solid tumours although their underlying causes of chromosomal instability in tumours remain obscure. In this paper we show how Artificial Immune System (AIS) paradigm can be successfully employed in the elucidation of biological dynamics of cancerous processes using a novel fuzzy rule induction system for data mining (IFRAIS) 1 of aCGH data. Competitive results have been obtained using IFRAIS. A biological interpretation of the results carried out using Gene Ontology is currently under investigation.

1. Introduction

Breast cancer (BC) is the most commonly diagnosed female cancer and, for this reason, even the most extensively investigated in terms of histopathology, immunochemistry and familial history. During recent studies an interesting trend has been observed: frequent losses and amplifications of chromosomal regions are concentrated in relatively small zones. This aspect pushed the interest for gene copy number studies. Each human being should have two copies of the same gene, however, due to biophysical reasons, it is possible that DNA strand breakage prevents DNA replication procedures to be correctly accomplished. If similar phenomena involve duplication of oncogenes or loss of oncosuppressor or DNA-repairing genes, cancerous processes can result to be activated. Changes in copy numbers of genes such as ERBB2 and c-MYC have been extensively documented in breast cancer and are localized in model cell lines [3–6]. Amplified (and overexpressed) genes are prime therapeutic targets as for example, the use of the drug trastuzumab against ERBB2 has been shown to improve breast cancer survival rates alone or in combination with other treatments [7–9]. On the other hand, aCGH proved to be a valuable tool in the investigation of biological dynamics underlying cancer [10]. Array CGH, namely aCGH, have greatly improved the resolution of this technology, enabling the detection of segmental copy losses and gains [11,12]. In this work we propose an aCGH approach to AI based investigation of familiali ty in BC patients. It is known, in fact, that familiality plays an important role in the explanati on of breast cancer cases under determined conditions. Here we show how a novel computational paradigm can be used in order to highlight inner characteristics of the disease: Induction of Fuzzy Rules with Artificial Immune Systems (IFRAIS) is a new algorithm developed to extract fuzzy classification rules from data [1]. IFRAIS' results are being validated using statistical driven approaches using Gene Ontology through GO Miner [15]. With this study we tried to show how IFRAIS can be employed in the design of experimental pipeline in disease processes investigation

*Corresponding author.

1448-5869/08/$17.00 © 2008 – IOS Press. All rights reserved

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.

2. Materials and methods

2.1. Specimens, data acquisition and data preprocessing

Samples' collection and data preprocessing steps have been carried out following protocols described elsewhere [10].

2.2. Algorithms

Common approaches to data mining in genomic datasets are mainly based on clustering techniques. However, the interpretation of the resulting clusters can reveal to be rather difficult due to the high amount of information that needs to be filtered in order to obtain interpretable information. Moreover there is 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" model of the system being developed (Artificial Neural Networks or Support Vector Machines, for example). Similar kinds 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 to judge its performances. This is a quite interesting perspective: underlying their interest there is the desire of gaining a deeper knowledge of the biological process by interpreting 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 to address these requests. The second set of approaches, then, gives a deeper insight into the problem adding to the prediction a clear description of how the prediction was made. Such clear descriptions can be represented by IF-THEN classification rules and the process of rule extraction from a dataset is called rule induction. Several algorithms have been proposed for accomplishing the rule induction task, being C4.5, probably, the most famous one. Moreover, in the recent years, research groups have tried to take advantage of soft computing and bio-inspired paradigms to develop more powerful and versatile data mining systems. The use of similar systems give rise to a new interest for such paradigms; these reasons suggested to include novel bio-inspired data mining systems in our comparative study. In the next subsection we will give a brief overview of IFRAIS; a detailed description of the ideas and concepts behind IFRAIS will be given.

2.3. Induction of fuzzy rules with artificial immune systems

The most important characteristic of IFRAIS (Induction of Fuzzy Rules with an Artificial Immune System) is that it discovers fuzzy classification rules [1]. This fuzzy format to rules is naturally comprehensible to human being. Nowadays, comprehensible knowledge is essential in real-world data mining problems (e.g. in bioinformatics). Hence, IFRAIS's discovered knowledge is not a "black-box" In essence, IFRAIS evolves a population of antibodies, where each antibody (Ab) represents the antecedent (the "IF part") of a fuzzy classification rule. Each antigen (Ag) represents an example (record, or case). The rule antecedent is formed by a conjunction of conditions (e.g., IF BAC1 is HIGH and BAC2 is LOW). Each attribute can be either continuous (real-valued, e.g. Salary) or categorical (nominal, e.g. Gender), as usual in data mining. Categorical attributes are inherently crisp, but continuous attributes are fuzzified by using a set of three linguistic terms (low, medium, high). Linguistic terms are represented by triangular membership functions, for the sake of simplicity. Each Ab is encoded by a string with n genes, where n is the number of attributes. Each gene $i, i = 1, \ldots, n$ consists of two elements: (a) a value V_{ij} specifying the value (or linguistic term) of the i^{th} attribute in the j^{th} rule condition; and (b) a boolean flag B_i indicating whether or not the i^{th} condition occurs in the classification rule decoded from the Ab. Hence, although all Abs have the same genotype length, different antibodies represent rules with different number of conditions in their antecedent – subject to the restriction that each decoded rule has at least one condition in its antecedent. This flexibility is essential in data mining, where the optimal number of conditions in each rule is unknown a priori. The rule consequents (predicted classes) are not evolved by the AIS. Rather, all the antibodies of a given AIS run are associated with the same rule consequent, so that the algorithm is run multiple times to discover

rules predicting different classes. The IFRAIS algorithm is based on two main procedures: the Sequential Covering (SC) and the Rule Evolution (RE).

2.4. Inducing rules from data

Each run of the algorithm discovers one fuzzy classification rule, so that the algorithm has to be run multiple times to discover multiple rules. This is obtained by using the SC procedure (often used in rule induction), as follows. The SC procedure starts with an empty set of discovered rules. Then it performs a loop over the classes. For each class, the algorithm will be run as many times as necessary to discover rules covering all or almost all the examples belonging to that class. More precisely, for each class the procedure initialises variable TS with the set of all examples in the training set, and then calls the AIS algorithm to discover a classification rule predicting the current class. The AIS returns the best evolved rule, which is added to the set of discovered rules. Next, the SC procedure removes from TS the examples that are correctly covered by the discovered rule, i.e. the examples that satisfy the rule antecedent and have the class predicted by the rule. Then the AIS algorithm is called again, to discover a rule from the reduced training set, and so on. This iterative process is repeated until the number of uncovered examples of the current class is smaller than a small threshold, called MaxUncovExamp (maximum number of uncovered examples). This avoids that the AIS tries to discover a rule covering a very small number of examples, in which case the rule would not be statistically reliable. This process is repeated for all the classes, producing a set of fuzzy classification rules covering almost all training examples. At the end of this training phase, the fitness of all rules is recomputed by considering the entire training set, in order to have a better estimate of rule quality to be used in the classification of test examples. The RE procedure starts by randomly creating an initial population of Ab. For each rule (Ab), the system prunes the rule – using the rule pruning procedure proposed by [35] to remove irrelevant conditions. Rule pruning has a twofold motivation: reducing the overfitting of the rules to the data and improving the simplicity (comprehensibility) of the rules [36]. Next, it computes fitness for each Ab. The fitness of an antibody Ab, denoted by fit(Ab), is given by Eq. (1):

$$fit(Ab) = \frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP} \quad (1)$$

In most projects using this function the discovered rules are crisp, whereas in IFRAIS the rules are fuzzy. Hence, the computation of the TP, FN, TN and FP involves, for each example, measuring the degree of affinity (fuzzy matching) between the example (Ag) and the rule (Ag). This is computed by applying the standard aggregation fuzzy operator min given by Eq. (2):

$$Affin(Ab, Ag) = \min_{i=1}^{n} (\mu_{Ab_i}(Ag_i)) \quad (2)$$

where $\mu_{Ab_i}(Ag_i)$ denotes the degree to which the corresponding attribute value of the example belongs to the fuzzy set associated with the ith rule condition, n is the number of conditions in the rule antecedent, and min is the minimum operator. An example satisfies a rule if the degree of affinity between the rule and the example is greater than an activation threshold, i.e., if $Affin(Ab, Ag) > L$.

For each antibody to be cloned the algorithm produces c clones. The value of c is proportional to the fitness of the antibody. More precisely, c increases linearly with the antibody fitness when $0 < Fit(Ab) < 0.5$, and any antibody with a fitness greater than or equal to 0.5 will have *MAXNUMCLONES* clones.

Next, each of the just-produced clones undergoes a process of hypermutation, where the mutation rate is inversely proportional to the clone's fitness (i.e., the fitness of its "parent" antibody). More precisely, the mutation rate for a given clone cl, denoted mut_rate(cl), is given by Eq. (3):

$$mute_{\text{rate}}(cl) = \alpha + (\beta - \alpha) \cdot (1 - fit(cl)) \quad (3)$$

where α and β are the smallest and greatest possible mutation rates, respectively, and $fit(cl)$ is the normalised fitness of clone cl. These numbers represent the probability that each gene (rule condition) will undergo mutation. Once a clone has undergone hypermutation, its corresponding rule antecedent is pruned by using the previously- explained rule pruning procedure. Finally, the fitness of the clone is recomputed, using the current TrainSet. In the next step the T-worst fitness antibodies in the current population (not including the clones created by the clonal selection procedure) are replaced by the T best-fitness clones out of all clones produced by the clonal selection procedure. Finally, the RE procedure returns, to the caller SC procedure, the best evolved rule, which will then be added to the set of discovered rules by the caller procedure.

Table 1
Overview of the results

	Accuracy	K-Statistic
IFRAIS	96.69%	0.848

3. Results

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. IFRAIS related tests were repeated 100 times to account for intrinsic variabili ty 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. The strategy for training and validation selected was the K-Fold cross-validati on with $K = 5$. The results are shown in Table 1.

Results returned by the experiments carried out show a quite interesting situation IFRAIS reaches a good absolute performance level. The global level of accuracy reached by the system nears the 97%; a quite competitive result indeed, even if we consider that algorithms like J48 (C4.5 evolution) is not able to go beyond the 94.34% of accuracy (results in the supplementary material). The rules extracted by IFRAIS are reported in supplementary material. Even if the statistical strength of the results can be considered a good index, confidence with results grows strongly with the understanding of the mechanism underlying decisions. For these reasons and the nature of the research we are carrying out validation of the results using a knowledge driven approach. We are employing Gene Ontology (GO) and BioCarta to discover interesting patterns in the rules extracted by the systems.

4. Discussion and cues for further research

In this work we presented a study of a novel rule induction system. This system has been used in mining the genomic data to extract useful knowledge in experimental oncology, validating the 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. For these reasons we tried to estimate how a novel approach to the target classification problem performed. We focused our research on systems generating fuzzy rules because of specific requests experts made in terms of system behaviour interpretability and reliability estimation. Fuzzy rules, in fact, can help handling a relative uncertainty about measurement that is particularly significant in the microarray scanning context. IFRAIS obtained good absolute and relative performances, similar to the performances of J48, an evolution of C4.5 algorithm. The global level of accuracy and Kappa-Statistic calculated over these systems allows us to be moderately confident about the rules generated and their coverage. Biological interpretation of the results, are being carried out using GO and BioCarta, however preliminary results show strong enrichment in GO Terms like the ones involved in FLJ and G-proteins that have been extensively documented in literature [2].

Moreover several interesting pathways and genes have been highlighted whose function and role in breast cancer inheritance mechanisms is currently under investigation (e.g. immune system related) [37]. We can conclude that novel biologically-inspired data mining techniques seem to be competitive interesting tools in cancer research. However the full understating of the underlying dynamics in cancer settlement and progression still remains a primary objective. In this context novel hints for research thought to improve IFRAIS have been formulated and include: (a) testing with others membership function formats, i.e. Gaussian; (b) the system could automatically determine the number of linguistic terms for each continuous attribute, rather than just using a fixed number as in the current version; (c) aggregating others immune principles, i.e. immune network; (d) remodelling it to cope with a multi-label classification problem. Moreover further studies are being carried out to optimise the number of features to be included 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 IFRAIS classifier and the use of fuzzy rules to model biological mechanisms underlying a complex process like breast cancer, which is an insidious disease whose understanding is slowly being incorporated within the expanding boundaries of our knowledge.

References

[1] R.T. Alves et al., An artificial immune system for fuzzy-rule induction in data mining. *Proc. of the 8th International Conference on Parallel Problem Solving from Nature* (*PPSN*), LNCS 3242, 2004. 1011–1020.

[2] D.G. Albertson, Profiling breast cancer by array CGH, *Breast Cancer Res Treat* **78** (2003) 289–298.

[3] O.P. Kallioniemi et al., ERBB2 amplification in breast cancer analysed by fluorescence in situ hybridization, *Proc Natl Acad Sci USA* **89** (1992), 5321–5325.
[4] M. Shimada et al., Detection of Her2/neu, c-MYC and ZNF217 gene amplifica tion during breast cancer progression using fluorescence in situ hybridization, *Oncol Rep* **13** (2005), 633–641.
[5] T.A. Jarvinen et al., Amplification and deletion of topoisomerase IIalpha associate with ErbB-2 amplification and affect sensitivity to topoisomerase II inhibitor doxorubicin in breast cancer, *Am J Pathol* **156** (2000), 839–847.
[6] M. Lacroix, Relevance of breast cancer cell lines as models for breast tumours: an update, *Breast Cancer Res Treat* **83** (2004), 249–289.
[7] L.A. Emens, Trastuzumab in breast cancer, *Oncology* (*Williston Park*) **18** (2004), 1117–1128.
[8] J. Baselga, Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials, *Oncology* **61**(Suppl 2) (2001), 14–21.
[9] C.L. Vogel et al., First-line Herceptin monotherapy in metastatic breast cancer, *Oncology* **61**(Suppl 2) (2001), 37–42.
[10] F. Menolascina et al., *Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective*, CIBCB 2007.
[11] J.J. Davies, Array CGH technologies and their applications to cancer genomes, *Chromosome Res* **13** (2005), 237–248.
[12] D. Pinkel et al., Array comparative genomic hybridization and its applications in cancer, *Nat Genet* **37**(Suppl) (2005), S11–S17.
[13] A. Chan et al., A New Ant Colony Algorithm for Multi-Label Classification with Applications in Bioinformatics, *Proceedings of GECCO 2006,* ACM Press, 2006, 27–34.
[14] M.H. Marghny et al., Extracting Logical Classification Rules with Gene Expression Programming: Microarray Case Study. In *Proc. of the International Conference on Artificial Intelligence and Machine Learning,* AIML 2005, Cairo, Egypt, 2005.
[15] B.R. Zeeberg et al., GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data, *Genome Biology* **4**(4) (2003), R28.
[16] D.G. Albertson et al., Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene, *Nature Genetics* **25** (2000), 144–146.
[17] D. Pinkel et al., High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays, 17 *Nature Genetics* **20** (1998), 207–211.
[18] S. Solinas-Toldo et al., Matrix-based comparative genomichybridization: biochips toscreen for genomic imbalances, *Genes Chromosomes Cancer* **20** (1997), 399–407.
[19] D. Pinkel et al., Quantitative high resolution analysis of DNA-copy number variation in breastcancer using comparative genomic hybridization to DNA microarrays, *Nat Genet* **20** (1998), 207–211.
[20] J.R. Pollack et al., Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *NatGenet* **23** (1999), 41–46.
[21] A.M. Snijders et al., Assembly of microarrays for genome-wide measurement of DNA copy number by CGH. *NatGenet* **29** (2001), 263–264. SPOT: http://cancer.ucsf.edu/array/analysis/index.php.
[22] M.S.B. Sehgal et al., Collateral Missing Value Imputation: a new robust missing value estimation algorithm for microarray data, *Bioinformatics* **21**(10) (2005), 2417–2423.
[23] I.S. Kohane et al., *Microarrays for an Integrative Genomics*, Cambridge, MA: MIT Press, 2003.
[24] L. Ein-Dor et al., Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21** (2006), 171–178.
[25] Gopalakrishnan et al., *Rule Learning for Disease-specific Biomarker Discovery from Clinical Proteomic Mass Spectra*, KDLL 2006.
[26] Li et al., Discovery of significant rules for classifying cancer diagnosis data, *Bioinformatics* **2** (2003), 93–102.
[27] A. Stanikov et al., A comprehensive evaluation of multicategory classifica tion methods for microarray gene expression cancer diagnosis, *Bioinformatics Advance Access* (September 16, 2004).
[28] T. Golub et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science* **286** (October 15, 1999).
[29] J.R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA, 1993.
[30] R.S. Parpinelli et al., Data mining with an ant colony optimization algorithm, *IEEE Transactions on Evolutionary Computation, special issue on Ant Colony Algorithms* **6**(4) (2002), 321–332.
[31] C. Ferreira, Gene Expression Programming in Problem Solving, in: *Soft Computing and Industry: Recent Applications*, R. Roy, M. Köppen, S. Ovaska, T. Furuhashi and F. Hoffmann, eds, Springer-Verlag, 2002, pp. 635–654.
[32] L.N. de Castro et al., Artificial Immune Systems: A New Computation Intellige nce Approach. Springer-Verlag, Berlin Zadeh L.A., *Fuzzy Sets Inform Control* **9** (1965), 338–352.
[33] W. Pedrycz et al., An Introduction to Fuzzy Sets, *Analysis and Design* (1998), MIT Press, Cambridge.
[34] H. Ishibuchi et al., Effect of Rule Weights in Fuzzy Rule-based Classifica tion Systems, *IEEE T Fuzzy Syst* **9**(4) (2001), 506–515.
[35] D.R. Carvalho et al., A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining, in: *Proc of GECCO 2000,* 2000, 1061–1068.
[36] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Mateo, 2 edition, 2005.
[37] K.E. de Visser et al., Paradoxical Roles of the Immune System During Cancer Development, Nature Reviews, 1 Supplementary material, 2006.