BMC Systems Biology

Poster presentation

Data mining techniques in a CGH-based breast cancer subtype profiling: an immune perspective with comparative study

Filippo Menolascina*[1,2], Stefania Tommasi[1], Patrizia Chiarappa[1], Vitoantonio Bevilacqua[2], Giuseppe Mastronardi[2] and Angelo Paradiso[1]

Address: [1]Clinical and Experimental Oncology Laboratory, National Cancer Institute, 70126 Bari, Italy and [2]Department of Electronics and Electrical Engineering, Polytechnic of Bari, 70126 Bari, Italy

Email: Filippo Menolascina* - f.menolascina@ieee.org

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):P70 doi:10.1186/1752-0509-1-S1-P70

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Background

Array Comparative Genomic Hybridization has been successfully used in post-genomic cancer research studies [1,2]. In particular this technology has been developed in order to monitor gene copy number changes in whole DNA. Results returned by similar screening techniques are in the form of microarray high dimensional data; the data complexity naturally requires computational analysis tools to extract reliable knowledge from the data. The discovery of such knowledge can then ease the difficulty of translating the complex raw data into relevant and clinically useful diagnostic or prognostic rules.

We applied data mining techniques and novel Artificial Intelligence immune inspired algorithms in order to analyse a dataset composed by 119 breast cancer samples divided in ER+ and ER- sets. The main objective of this analysis was to find genes involved in the activation of Estrogen Receptor. Several approaches have been exploited and their results are compared. Both predictive power of classifiers and derived biological interpretation are reported and discussed. Promising results have been showed by C4.5 derived classifier and immune based approaches that pushed for further research in employment of similar systems in this field.

Several classifier have been developed in order to compare the ability of different approaches in classifying the data according to the binary discrimination ER+/ER-. The dataset under investigation was composed as follows by 33 ER- cases and 86 ER+ cases-. Interesting rules have been extracted from the dataset under investigation. These and other comments about the results are discussed in the final section of this abstract. Accuracy and Kappa-Statistic are reported in table 1 and a graphical representation of the same results is given in table 1. Each system has been trained on the 66% of the 119 cases of the dataset, leaving 41 cases for validation purposes. The results reported refer to the validation set.

All of the systems under investigation in this analysis showed a quite high accuracy. However as it can be seen in Table 1, there's a interesting separation between the Boosting and Bagging based systems and the other approaches. In particular J48 showed the best absolute accuracy although JRip, AIRS and Immunos returned comparable results. Although the expressive power of J48 systems is a well established characteristic of these kind of approaches, the potentialities of immune based systems is currently an open field of research. Then, these results, seem to confirm the promising aptitude of immune inspired paradigms to stand at the basis of accurate classifiers for data mining tasks in bioinformatics.

Table 1: Results returned by each system

	Bagging	AdaboostM1	Logit	MultiBoost	J48	JRip	AIRS	Immunos	CSCA
Accuracy	82.37%	82.93%	85.37%	85.37%	90.24%	87.80%	87.80%	87.80%	82.93%
K-Stat	0.502	0.393	0.541	0.502	0.694	0.602	0.602	0.632	0.393

The samples analysed in this study were acquired and collected as described elsewhere [3]. One hundred and nineteen cases, each of which composed by 2424 features composed the raw dataset. Data pre-processing techniques were employed in order to reduce the impact of noise and artifacts derived from data acquisition, in particular gene filtering and raw value normalization has been applied. The obtained dataset has been splitted in two subsets (ER+ and ER- classes) and a set of class separability has been studied using Student T-test and Entropy criteria. A comprehensive ranking of the genes best representing discriminant features has been obtained computing a consensus estimate of the position in the previous two classifications. The first 100 genes in this new ranking were used for further analyses. A new dataset has been built on these new data, counting 119 cases and 100 observations for each case. Several different classifier having different peculiar strength points have been built with the only objective of creating a common platform by which a coherent comparative study could be set. For these reasons common Bagging and Boosting approaches have been used together with typical tree classifier and immune inspired ones. In the last category lay all those systems that use paradigms imported from human immunology in order to reproduce adaptive behaviours that allow our immune system to reject the attacks of pathogens and potentially harmful molecules.

In this study we have compared the performances of all of Artificial Immune Recognition System (AIRS) [4], Clonalg and CSCA [5] and Immunos [6] systems with well established data mining tools like tree classifiers (J48 and JRip) and meta-learners (Bagging, AdaBoostM1 (Boosting with Decision Stump), Logit (performing logistic regression), and MultiBoost (and extension of AdaBoostM1)). In addition we used Kappa statistic as a measure of the agreement between predicted and observed categorization of the dataset under investigation, while correcting for agreement that occurs by chance.

In this study the performances of different data mining and AI based systems for high throughput data classification have been compared. The results put in evidence an interesting trend: tree classifier J48, an extension of the C4.5 system, showed the best accuracy among all of the systems taken into account for this study. Although showing a high absolute accuracy, this kind of classifier is also able to maintain a good expressive power by returning trees that can be easily translated in rules.

These rules can be further interpreted by a human expert or reintroduced in a knowledge driven validation pipeline that takes advantage of tools like, for example, Gene Ontology [6]. The results showed interesting trends, indeed. Artificial Immune Systems based classifier, in fact, returned results, in terms of accuracy and kappa statistic quite comparable with the ones that characterized best performing tree classifiers. These results seem to encourage further studies on the employment of such systems in these context; AIS systems seem to be their ease in context characterized by high dimensional data and complex information distribution. For these reasons our laboratory is now trying to repeat the same analysis, this time in order to classify familial and sporadic breast cancers.

References

1. Albertson DG: **Profiling breast cancer by array CGH.** *Breast Cancer Res Treat* 2003, **78:**289-298. doi: 10.1023/A: 1023025506386.
2. 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 USA* 2002, **99:**12963-12968. doi: 10.1073/pnas.162471999.
3. Menolascina F, Tommasi S, Fedele V, Paradiso A, Mastronardi G, Bevilacqua V: **Hybrid Intelligent Data Mining Techniques and Array CGH in Breast Cancer Profiling.** in press.
4. Timmis J, Boggess J: **Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm.** *Genetic Programming and Evolvable Machines* 2004, **5:**291-317.
5. de Castro LN, Von Zuben FJ: **The clonal selection algorithm with engineering applications.** *Artificial Immune Systems* 2000, **8:**36-39.
6. Carter : **The Immune System as a Model for Pattern Recognition and Classification.** *J Am Med Inform Assoc* 2000, **7:**28-41.