Analysis of Cancer Microarray Data using Constructive Neural Networks and Genetic Algorithms

R. M. Luque-Baena, D. Urda, J.L. Subirats, L. Franco, and J.M. Jerez

Department of Computer Science, University of Málaga, Málaga, Spain;
{rmluque,durda,jlsubirats,lfranco,jja}@lcc.uma.es

Abstract. The analysis of microarray data typically involves a feature selection method in order to select the most relevant genes while at the same time maximizing the information content. This work presents a methodology that use the Welch t-test to filter the number of initial features embedded in two different frameworks to select the predictor genetic profile: genetic algorithm and stepwise forward selection approaches. The genetic algorithm strategy combines mutual information and classification models to predict cancer outcome. Furthermore, a constructive neural network model, C-Mantec, is applied providing reduced network architectures with competitive results in comparison to other classifiers. Six free-public cancer databases are used to test our approach.

Keywords: Microarray, Genetic algorithms, Constructive neural networks

1 Introduction

DNA microarray technology has been widely used in cancer studies for prediction of disease outcome [1]. It is a powerful platform successfully used for the analysis of gene expression in a wide variety of experimental studies [2]. However, due to the large number of features (in the order of thousands) and the small number of samples (mostly less than a hundred) in this kind of datasets, microarray data analysis face the "large-p-small-n" paradigm [3] also known as the curse of dimensionality. In this sense, the microarray data analysis usually involves a preprocessing step, which consists in the selection of features (genes) relevant for the classification step.

In this approach, a feature selection method based on genetic algorithms (GAs) and classification methods is proposed, focusing on constructive neural networks (CNNs), C-Mantec in particular, as a competitive choice for classification/prediction tasks. On one hand, GAs are well considered as suitable evolutionary strategies for feature selection in the literature. They are well adapted for problems with a large number of features [4], and are applied to different areas, from object detection [5] to gene detection in microarray data [6]. The use of non redundant features is sometimes preferable. Thus, our strategy incorporates a mutual information filter to minimize the correlation between the selected features, at the same time that it increases the classifier performance. On the other hand, CNNs have been proved to get similar classification results than traditional multi-layer perceptrons (MLP) or support vector machines (SVM), with the advantage that the architecture is dynamically estimated [7]. This is a critical factor in the wrapper selection methods combined with neural networks, because the subsets analyzed are from different sizes (or even the complexity of the features selected need projections in higher spaces), which implies that the use of the same architecture is not always appropriate.

Additionally, it is a novelty the application of these constructive networks in microarray analysis, generating a more compact design which can be more suitable in the bioinformatics field. Several comparison results are provided using other feature selection strategy (Stepwise Forward Selection method) and different classification techniques (LDA, SVM and Naive Bayes), in order to check the viability and suitability of the scheme proposed in this paper. These results are tested over six public cancer datasets (breast, colon, leukemia, lung, ovarian and prostate cancer) that are commonly used in the literature.

The remainder of this paper is organised as follows: Section 2 sets out the feature selection methodology of this approach describing the evolutionary strategy and the constructive classification model proposed as a competitive alternative to other well known classification models, and Section 3 shows the experimental results over several well-known public cancer databases. Finally, Section 4 concludes the article.

2 Feature Selection Framework

Feature selection techniques can be organized into three broad categories: filter, wrapper and embedded methods [8]. Filter methods use statistical properties of the variables to discard poorly descriptive features and are independent of the classifier. Wrapper methods are more computationally demanding than filter methods, where the subsets of features are evaluated within a classification algorithm with a measure of the goodness of a feature subset as the improvement criteria. Embedded methods are also classifier dependent, but they can be viewed as a search in the combined space of feature subsets and classifier models. Thus, it is not possible to replace one classifier with other different, since the feature selection and the classification method work as a whole.

This approach tries to achieve a two-fold objective; on the one hand, it is necessary to select the most relevant genes with a significant influence in the disorders which are being studied; on the other hand, good generalization rates in the prediction stage are essential to determine the probability of suffering from a specific condition. It is known that these approaches can improve the classification performance by discarding either irrelevant or redundant features.

2.1 Stepwise Forward Selection procedure

An exhaustive evaluation of all the possible subsets of n features involves a complexity of $\mathcal{O}(2^n)$ which becomes unfeasible for large values of n. In this sense, many heuristic algorithms have been proposed to reduce the computational complexity of wrapper algorithms. Stepwise forward procedures for feature selection analyze the inclusion of one or several features in order to improve the performance of the classification task. Thus, sequential forward selection [9] chooses the best variable in each iteration by minimizing the misclassification rate, and includes it in the final subset of features, starting with an empty set. The algorithm will continue to add variables until the resulting subset does not improve, in terms of an specific criteria.

2.2 Methodology approach

In this paper a methodology approach is presented based on GAs [5] and mutual information [10]. It can be viewed as a wrapper method with a combination of filter approaches (for removing feature redundancy), and classification methods. Since cancer databases provide a huge number of genes, a pre-selection step to reduce the number of variables is required. The Student´s t-test has been found more successful than other filter methods in terms of the quality of the features ranked [11]. Specifically, the Welch´s t-test [12], which is an adaptation of the previous one, is applied assuming the two classes (the patient has cancer or not) have unknown and unequal variances, because it is not advisable to use the basic form if we are unsure if the requirements of the test are satisfied [8]. A 5% of the total number of genes are retained (between 400 and 2000 genes, approximately, in the datasets selected), which will be the input of the genetic algorithm of the next section.

Evolutionary Strategy. GAs are a class of optimization procedure inspired by the biological mechanisms of reproduction. In this kind of optimization problems, a fitness function $f(\mathbf{x})$ should be maximized or minimized over a given space X of arbitrary dimension.

Encoding and Initial Population. A simple encoding scheme to represent as much as possible of the available information was employed, in which the chromosome is a string of bits whose length is determined by the total number of genes. Each variable is associated with one bit in the string. If the i^{th} bit is active (value 1), then the i^{th} gene is selected in the chromosome. Otherwise, a value of 0 indicates that the corresponding feature is ignored. In this way, each chromosome represents a different feature subset. Both, the active features and the number of them are generated randomly. In all the experiments, the population size of 100 individuals was used.

Selection, Crossover and Mutation. A selection strategy based on roulette wheel and uniform sampling was applied, while an elite count value of 10 (number of chromosomes which are retained in the next generation) was selected. Scattered crossover, in which each bit of the offspring is chosen randomly, was the choice for combining parents of the previous generation. The crossover rate was set to 0.8. In addition to that, a traditional mutation operator which flips a specific bit with a probability rate of 0.2 was considered. A modification which involves mutating a random number of bits between 1 and the number of active features of the individual is introduced. Since it was empirically verified that the best subsets include few features, this change avoids the increment on the number of active features in the last generations of the GA.

Fitness function. The fitness function assesses each chromosome in the population so that it may be ranked against all the other chromosomes. The main goal of feature subset selection is to use fewer features to achieve the same or better performance. Additionally, it has been found that the combination of features with low redundancy among them, i.e., by providing different information about the target class, and with a certain resemblance to the target class, can improve the performance rates [13]. Therefore, the fitness function should contain three terms: the misclassification error, the number of features selected and a redundancy measure among them. Datasets are splitted into training and testing sets in order to evaluate the generalization ability of the proposed chromosome.

Statistical techniques such as mutual information [10] give us an idea of the correlation between a pair of features. The mutual information between two continuous random variables x and y is given by

$$I(y,z) = \int\int p(y,z)\log\left(\frac{p(y,z)}{p(y)\,p(z)}\right) dy\,dz \tag{1}$$

where $p(y,z)$ is the joint probability density function of y and z, and $p(y)$ and $p(z)$ are the marginal probability density functions of y and z respectively. The mutual information is symmetric.

Moreover, it is non-negative, with a zero value indicating that the variables are independent. The more correlated two variables are, the greater their mutual information. Advantages of mutual information are that the dependency between variables is no longer restricted to be linear and it can handle nominal or discrete features. Although it is hard to compute for continuous data, the probability densities can be discretized using histograms, which are considered as good approximations [14]. A measure which incorporates the correlation of the features with the target class and penalizes the redundancy among the selected features is described as follows [13]:

$$corr(\mathbf{x}) = \frac{1}{t}\sum_{i=1}^{k}\sum_{j=i+1}^{k} I(x_j,x_i) - \frac{1}{k}\sum_{j=1}^{k} I(x_j,C) \tag{2}$$

where k is the number of features selected, C is the target class and t is the number of combinations between the pairs of the chromosome x analysed. Finally, the function to be minimised is represented as follows:

$$fitness(\mathbf{x}) = (1 - ACC(\mathbf{x})) + \lambda\frac{k}{\mathcal{N}} + \beta corr(\mathbf{x}) \tag{3}$$

where $fitness(\mathbf{x})$ is the fitness value of the feature subset represented by $\mathbf{x}$; $ACC(\mathbf{x})$ is the accuracy rate obtained by the classifier using the test set; $\mathcal{N}$ is the total number of extracted features; finally, $corr(\mathbf{x})$ defines the correlation among the features and the target class, with the aim of avoiding the redundancy in the feature vector (equation 2). The parameters λ and β can take values in the interval $(0,1)$ and were empirically chosen to 0.4 and 0.25, respectively.

Therefore, if two subsets achieve the same performance, while containing a different number of features, the subset with fewer features is preferred. We also stimulate the mixture of features less redundant among them, which is considered a good quality for classification tasks. Nevertheless, among the three terms, error, feature subset size, and correlation, the first one is our major concern.

C-Mantec algorithm. C-Mantec (Competitive Majority Network Trained by Error Correction) [7] is a novel neural network constructive algorithm that utilizes competition between neurons and a modified perceptron learning rule to build compact architectures with good prediction capabilities. The novelty of C-Mantec is that the neurons

Dataset	*#Genes*	*Samples*	*Class 0*	*Class 1*	*Data Proportion*
			(normal)	***(cancer)***	
Leukemia	7129	72	25	47	0.347
Lung	12533	181	150	31	0.829
Colon	2000	62	22	40	0.355
Breast2Class	4869	78	33	44	0.423
Ovarian	15154	253	91	162	0.360
Prostate	12600	102	50	52	0.490

Table 1. Information about the six databases analyzed.

compete for learning the new incoming data, and this process permits the creation of very compact neural architectures. At the single neuronal level, the algorithm uses the thermal perceptron rule, introduced by Marcus Frean in 1992 [15], that improves the convergence of the standard perceptron for non-linearly separable problems. C-Mantec, as a CNN algorithm, has in addition the advantage of generating online the topology of the network by adding new neurons during the training phase, resulting in faster training times and more compact architectures [16, 17]. Its network topology consists of a single hidden layer of thermal perceptrons that maps the information to an output neuron that uses a majority function.

3 Experimental Results

In this section, six free-public cancer databases[1] have been used to test our methodology. The information of each dataset is shown in Table 1. Two different comparison frameworks are raised. Thus, the GA approach is compared to the classical stepwise forward selection (SFS), where for each methodology several classification techniques are applied, namely: linear discriminant analysis (LDA), support vector machines (SVM), naive Bayes (NB) and the constructive neural network proposed (C-Mantec). With regard to the parameter configuration, both LDA and NB use the default parameters whereas SVM uses a radial basis function as kernel with a cost of $C = 10$. On the other hand, the following values $I_{max} = 10000$ $Fi_{temp} = 3$ and $G_{fact} = 0.2$ are empirically selected for C-Mantec. For each classifier, a holdout validation strategy is used by dividing the entire dataset on $60\% - 40\%$, the first one to train the model and the second to get the accuracy result. This training-testing procedure is launched 50 times varying the training and testing set to avoid the highly dependency of the evaluation.

The comparison results between the previous frameworks are observed in Table 2. By analysing the two feature selection methodologies, *Leukemia, Lung* and *Ovarian* databases are successfully analysed, with accuracy rates close to 100% regardless of the classifier applied. The complexity of the datasets *Breast2Class, Colon* and *Prostate* is a little higher, which implies that the SFS algorithm does not manage to obtain suitable rates with this number of genes. On the contrary, although the GA selects a higher

[1] http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html

		LDA		SVM		NB		C-MANTEC	
		mean±std	*#genes*	*mean±std*	*#genes*	*mean±std*	*#genes*	*mean±std*	*#genes*
SFS	**Leukemia**	97.784±2.65	4	98.491±1.81	4	97.964±2.25	3	98.579±2.83	5
	Lung	99.961±0.23	3	99.980±0.17	5	99.984±0.15	3	99.743±0.83	5
	Colon	87.942±6.43	4	87.842±5.43	4	86.540±5.80	4	87.407±6.66	5
	Breast2Class	81.524±6.08	4	82.783±5.57	3	78.532±5.84	3	76.900±9.02	7
	Ovarian	99.978±0.15	4	99.972±0.17	4	99.938±0.25	3	99.826±0.51	4
	Prostate	93.996±3.20	4	96.152±2.72	6	95.518±3.10	4	91.476±4.14	5
	Average	**93.531±3.12**	3.83	**94.203±2.65**	4.33	93.079±2.90	3.33	92.32±4.00	5.17
GA	**Leukemia**	99.943±0.06	5	99.829±0.19	4	99.786±0.10	5	99.363±0.19	7
	Lung	99.672±0.15	4	85.533±0.69	66	99.989±0.02	6	99.467±0.13	5
	Colon	94.733±0.53	8	71.850±1.72	65	94.817±0.65	21	96.203±0.52	12
	Breast2Class	98.747±0.17	20	96.213±0.52	16	92.668±0.44	43	94.245±0.46	20
	Ovarian	99.628±0.07	2	99.864±0.13	3	99.960±0.04	3	99.265±0.11	2
	Prostate	99.750±0.10	11	99.950±0.05	10	98.940±0.27	9	99.308±0.13	11
	Average	**98.746±0.18**	8.33	92.207±0.55	27.33	97.693±0.25	14.50	**97.975±0.25**	9.50

Table 2. Performance comparison among two different feature selection frameworks (SFS and GA) and four classifiers (LDA, SVM, NB and C-MANTEC) for each cancer microarray dataset. The results correspond to the best simulation for each database, by showing the accuracy of each classification method in the format of *mean±standard deviation* and the number of genes selected.

number of genes for these databases, it obtains a better accuracy rate. Thus, the improvement of the GA methodology with regard to the SFS selection approach is notable since the search space analysed is wider and the heuristic fitness function leads to the aim correctly. With the exception of the GA-SVM proposal for the *Lung* and *Colon* data where the results could improve, the remaining combinations between GA and

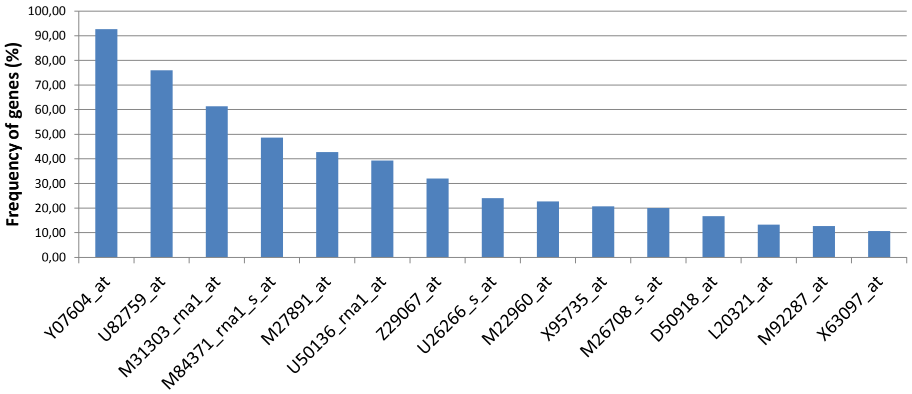

Fig. 1. Distribution of the most frequently selected genes (in 50 independent executions) by the GA-CMANTEC strategy in *Leukemia* dataset.

Index	Accession Num.	Gene Description	References
4951	*Y07604_at*	NME/NM23 nucleoside diphosphate kinase 4	[18–20]
3847	*U82759_at*	Homeo box A9	[21, 22, 20]
1928	*M31303_rna1_at*	Op 18	[21]
6225	*M84371_rna1_s_at*	CD19 Molecule	[23]
1882	*M27891_at*	CST3 Cystatin C	[21, 18, 20, 22]
3320	*U50136_rna1_at*	LTC4 synthase	[21, 20, 22, 23]
5107	*Z29067_at*	NIMA-related kinase 3	[20]
4847	*X95735_at*	Zyxin	[21, 18, 22, 23]
2354	*M92287_at*	CCND3 Cyclin D3	[21, 18, 22, 23]

Table 3. Some of the best genes ranked with GA-CMANTEC which appear in other studies in the literature.

any other classifier overcome in terms of classification accuracy the obtained ones for the SFS approach. Additionally, the robustness of the selected features is considerably higher in the GA (less standard deviation in the classification task on average), since this approach evaluates the best subset of features (chromosomes) several times in the whole process.

The strategy GA-CMANTEC has resulted quite suitable, obtaining the second best result on average after LDA over all the datasets analysed. Therefore, it is possible to conclude that constructive neural networks and, concretely, the C-Mantec classifier provides good classification skills with a competitive performance in comparison to the remaining alternatives in the bioinformatic field.

Focusing on the GA-CMANTEC approach, and as a brief biological analysis of the features selected, Figure 1 displays the most selected genes, after 50 independent executions, for the *Leukemia* database which is one of the most studied dataset in the literature. In order to check the coherence of the selection, Table 3 shows the best selected genes which also have been extracted in several related papers (last column of the table). It should be noted that the applied methodology is different from one paper to another. For instance, six of the nine genes are also reported in the list of the 50 most important genes (selected from 7129) suggested in [21]. Finally, a graph of the most frequent couples of genes is also presented in Figure 2, where stronger links are associated with thicker lines. It is possible to observe that the *Y07604_at* and *U82759_at* genes form a strong group which possibly provides a biological understanding of the leukemia disease.

4 Conclusions

In this work, a new methodology approach combining genetic algorithm with constructive neural networks has been proposed in order to predict cancer outcome. For six free-public cancer databases, we first select the most relevant features with a significant influence in the disorders studied comparing the GA with the SFS algorithm, testing the prediction accuracy using C-Mantec, LDA, SVM or NB as classifiers.

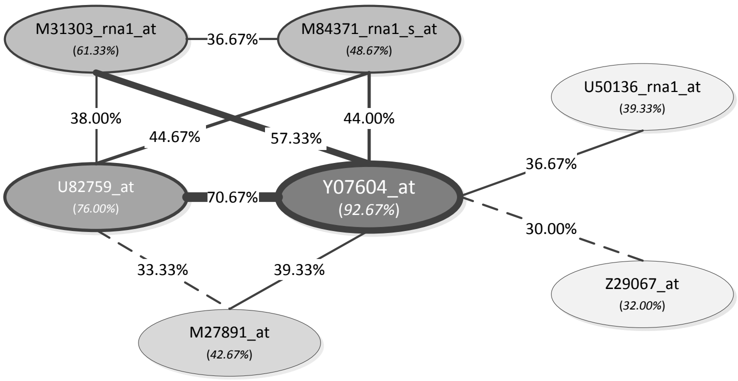

Fig. 2. Pairwise graph of the most frequently selected couples of genes in *Leukemia* dataset. It should be noted that stronger links correspond to thicker lines.

On average, the strategy based on the GA approach leads to better prediction rates, observing that these results are independent from the classifier used. Moreover, another advantage of the GA method is the lower variability for the accuracy. In addition, C-Mantec presents very competitive results in terms of prediction rate as well as selects some features of the Leukemia database that have also been published in the literature as the most significant ones related to this disease.

It could be interesting for future works to include biological information about the genes into the selection procedure according to the studied disease, instead of making this decision based only on the prediction rate.

Acknowledgements

The authors acknowledge support through grants TIN2010-16556 from MICINN-SPAIN and P08-TIC-04026 (Junta de Andalucía), all of which include FEDER funds.

References

1. Wei, J.S., Greer, B.T., Westermann, F., Steinberg, S.M., Son, C.G.: Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastom. Cancer Research **64** (2004) 6883 – 6891
2. Pellagatti, A., Vetrie, D., Langford, C.F., Gama, S., Eagleton, H., Wainscoat, J.S., Boultwood, J.: Gene expression profiling in polycythemia vera using cdna microarray technology. Cancer Research **63** (2003) 3940 – 3944
3. West, M.: Bayesian factor regression models in the “large p, small n” paradigm. Bayesian statistics **7**(2003) (2003) 723–732
4. Raymer, M., Punch, W., Goodman, E., Kuhn, L., Jain, A.: Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation **4**(2) (2000) 164–171
5. Sun, Z., Bebis, G., Miller, R.: Object detection using feature subset selection. Pattern Recognition **37**(11) (2004) 2165 – 2176

6. McLachlan, G., Bean, R., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**(3) (2002) 413–422
7. Subirats, J.L., Franco, L., Jerez, J.M.: C-mantec: A novel constructive neural network algorithm incorporating competition between neurons. Neural Networks **26** (2012) 130 – 140
8. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19) (2007) 2507–2517
9. Webb, A.R.: Statistical Pattern Recognition, 2nd Edition. Third edn. John Wiley & Sons (2011)
10. Guo, B., Nixon, M.: Gait feature subset selection by mutual information. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **39**(1) (2009) 36 –46
11. Huerta, E.B., Duval, B., Hao, J.K.: A hybrid lda and genetic algorithm for gene selection and classification of microarray data. Neurocomputing **73** (2010) 2375 – 2383
12. Welch, B.L.: The generalization of student´s problem when several different population variances are involved. Biometrika **34**(1-2) (1947) 28–35
13. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8) (2005) 1226 – 1238
14. Moddemeijer, R.: On estimation of entropy and mutual information of continuous distributions. Signal Processing **16**(3) (1989) 233 – 246
15. Frean, M.: A “thermal” perceptron learning rule. Neural Comput. **4**(6) (1992) 946–957
16. García-Pedrajas, N., Ortiz-Boyer, D.: A cooperative constructive method for neural networks for pattern recognition. Pattern Recogn. **40**(1) (2007) 80–98
17. Subirats, J.L., Jerez, J.M., Franco, L.: A new decomposition algorithm for threshold synthesis and generalization of boolean functions. IEEE Transactions on Circuits and Systems **1**(55) (2008) 3188–3196
18. Yang, P., Zhou, B.B., Zhang, Z., Zomaya, A.Y.: A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. BMC Bioinformatics **11**(1) (2010)
19. García-Nieto, J., Alba, E., Jourdan, L., Talbi, E.: Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. Information Processing Letters **109**(16) (2009) 887 – 896
20. Krishnapuram, B., Carin, L., Hartemink, A.: Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data. Journal of Computational Biology **11**(2-3) (2004) 227–242
21. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286**(5439) (OCT 15 1999) 531–537
22. Chen, Z., Li, J., Wei, L.: A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artificial Intelligence in Medicine **41**(2) (2007) 161 – 175
23. Momin, B., Mitra, S., Gupta, R.: Reduct generation and classification of gene expression data. In: Hybrid Information Technology, 2006. ICHIT ’06. International Conference on. Volume 1. (2006) 699 – 708