


# An Insight on the ‘Large $G$ , Small $n$ ’ Problem in Gene-Expression Microarray Classification

V. García<sup>1</sup>, J.S. Sánchez<sup>2</sup> , L. Cleofas-Sánchez<sup>3</sup>, H.J. Ochoa-Domínguez<sup>4</sup>,  
and F. López-Orozco<sup>1</sup>

<sup>1</sup> Multidisciplinary University Division, Universidad Autónoma de Ciudad Juárez,  
Ciudad Juárez, Chihuahua, Mexico

<sup>2</sup> Department of Computer Languages and Systems,  
Institute of New Imaging Technologies, Universitat Jaume I,  
Castelló de la Plana, Spain

[sanchez@uji.es](mailto:sanchez@uji.es)

<sup>3</sup> National Institute of Genomic Medicine, Ciudad de México, D.F., Mexico

<sup>4</sup> Department of Electrical and Computer Engineering,  
Universidad Autónoma de Ciudad Juárez,  
Ciudad Juárez, Chihuahua, Mexico

**Abstract.** This paper analyzes the effect of the high-dimensional, low-sample size problem in cancer classification using gene-expression microarrays. Here the two key questions addressed are: (i) What is the percentage of genes that can ensure highly accurate classification?, and (ii) Does this percentage differ from one classifier to another? Both these issues are investigated by developing a pool of experiments with two gene ranking algorithms, five classifiers and four DNA microarray databases.

**Keywords:** DNA microarray · Gene expression · Feature ranking · Cancer classification

## 1 Introduction

Conventional methods for cancer classification rely on a variety of morphological, clinical and molecular variables, but they exhibit several limitations that make difficult an accurate diagnosis. The rapid development of high-throughput biotechnologies such as DNA microarray analysis allow to record and monitor the expression levels of thousands of genes simultaneously from a few samples [8], which has attracted the attention of scientists for its application in basic and translational cancer research [5, 14, 15, 18]. Many studies utilizing DNA microarrays have been directed to (i) distinguish between cancerous and non-cancerous tissue samples, (ii) classify different types or subtypes of tumors, and (iii) predict the response to a particular therapeutic drug and/or the risk of relapse.

Cancer classification using microarrays, which focuses on predicting the class of a new sample based on its expression profile, poses two major challenges. First, the gene-expression data are characterized by the so-called ‘large  $G$ ,

small  $n'$  problem, that is, the number of genes ( $G$ ) heavily exceeds the sample size ( $n$ ). And second, most genes are irrelevant to discriminate samples of different types [6]. These issues may increase the complexity of the prediction problem, degrade the generalization ability of classifiers and hinder the understanding of the relationships between the genes and the tissue samples [4, 19]. Under these circumstances, feature selection plays a very important role in cancer classification because it can alleviate (minimize) the effects of both those problems.

A particularly popular approach to feature selection using DNA microarrays is gene ranking [9, 13, 17, 20]. Gene ranking methods are filters that encompass some scoring function to quantify how much more statistically significant each gene is than the others [7], and as a result they rank genes in decreasing order of the estimated scores under the assumption that the top-ranked genes correspond to the most informative (or differentially expressed) ones.

The question the present study intends to answer is how the ‘large  $G$ , small  $n$ ’ problem affects the classification performance using gene-expression microarrays. In particular, this paper examines the impact of high-dimensional biological data on several standard classifiers. To this end, two feature ranking algorithms are applied to select a percentage of the top-ranked genes, which are further used to classify new tissue samples and record the performance of classifiers in terms of both overall accuracy and false-negative rate.

## 2 Gene Ranking Algorithms

Some well-established gene ranking strategies include  $t$ -test, information-theoretic measures, symmetric uncertainty, correlation coefficient,  $\chi^2$ -statistic and ReliefF, among others. In this section, the two feature ranking methods used in the experiments are briefly described.

### 2.1 ReliefF

The basic idea of the ReliefF algorithm [12, 16] lies on adjusting the weights of a vector  $W = [w(1), w(2), \dots, w(G)]$  to give more relevance to features that better discriminate the samples from neighbors of different class.

It randomly picks out a sample  $x$  and searches for  $k$  nearest neighbors of the same class (hits,  $h_i$ ) and  $k$  nearest neighbors from each of the different classes (misses,  $m_i$ ). If  $x$  and  $h_i$  have different values on feature  $f$ , then the weight  $w(f)$  is decreased because it is interpreted as a bad property of this feature. In contrast, if  $x$  and  $m_i$  have different values on the feature  $f$ , then  $w(f)$  is increased. This process is repeated  $t$  times, updating the values of the weight vector  $W$  as follows

$$w(f) = w(f) - \frac{\sum_{i=1}^k \text{dist}(f, x, h_i)}{t \cdot k} \quad (1)$$

$$+ \sum_{c \neq \text{class}(x)} \frac{P(c)}{1 - P(\text{class}(x))} \cdot \frac{\sum_{i=1}^k \text{dist}(f, x, m_i)}{t \cdot k}$$

where  $P(c)$  is the prior probability of class  $c$ ,  $P(class(x))$  denotes the probability for the class of  $x$ , and  $dist(f, x, m_i)$  represents the absolute distance between samples  $x$  and  $m_i$  in the feature  $f$ .

### 2.2 Gain Ratio

The Gain ratio is an extension of information gain in order to overcome the biased behavior of selecting the features with the largest number of values. Let  $X$  be a set of  $n$  samples that belong to  $C$  distinct classes and let  $n_i$  be the number of samples in class  $i$ . The entropy of any subset can be calculated using the following formula

$$H(X) = - \sum_{i=1}^C ((n_i/n) \cdot \log(n_i/n)) \tag{2}$$

To find the information gain of feature  $f$ , one has to sum the entropy for each value  $f_j$  ( $j = 1, \dots, v$ ) of the feature:

$$H(X|f) = \sum_{j=1}^v ((|f_j|/n) \cdot H(X|f = f_j)) \tag{3}$$

where  $H(X|f = f_j)$  is the entropy calculated relative to the subset of instances that have a value of  $f_j$  for feature  $f$ .

The information gain of a feature is measured by the reduction in entropy as  $IG(f) = H(X) - H(X|f)$ . The greater the decrease in entropy when considering feature  $f$  individually, the more significant this is for prediction.

In general, a feature will be most useful when maximizing the information gain while simultaneously minimizing the number of feature values. Then the intrinsic value of a feature  $f$  can be computed as:

$$IV(f) = - \sum_{i=1}^v ((|f_i|/n) \cdot \log(|f_i|/n)) \tag{4}$$

Thus the Gain ratio of  $f$  is defined as

$$Gain\ ratio(f) = \frac{IG(f)}{IV(f)} = \frac{H(X) - H(X|f)}{H(f)} \tag{5}$$

## 3 Databases and Experimental Setting

We conducted a series of experiments on a collection of publicly available microarray cancer data sets taken from the Kent Ridge Biomedical Data Set Repository (<http://datam.i2r.a-star.edu.sg/datasets/krbd>). Table 1 summarizes the main characteristics of these data sets, including the number of genes (features), the number of tissue samples, and the size of the positive and negative classes.

**Table 1.** Characteristics of the gene-expression microarray data sets.

	#Genes	#Samples	Positive–Negative
Breast	24481	97	Relapse (46)–Non-relapse (51)
CNS	7129	60	Failure (39)–Survivor (21)
Colon	2000	62	Tumor (22)–Normal (40)
Prostate	12600	136	Tumor (77)–Normal (59)

The 5-fold cross-validation method was adopted for the experimental design because it appears to be the best estimator of classification performance compared to other methods, such as bootstrap with a high computational cost or re-substitution with a biased behavior [1].

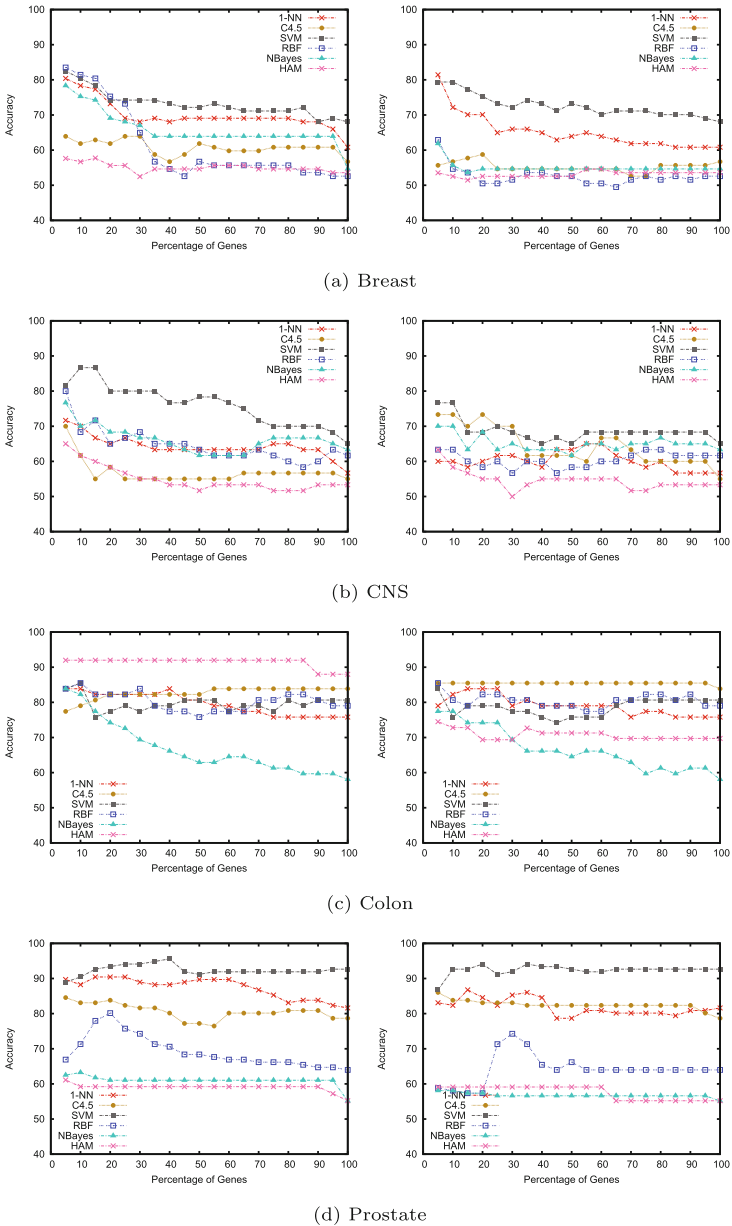
We focused our study on the ReliefF and Gain ratio feature ranking algorithms and five classification models: the nearest neighbor rule (1-NN), a support vector machine (SVM) using a linear kernel function with the soft-margin constant  $C = 1.0$  and a tolerance of 0.001, the C4.5 decision tree, the naive Bayes (NBayes) classifier, the radial basis function neural network (RBF) with the K-means clustering to provide the basis functions, and a hybrid associative memory (HAM) with translation of the coordinate axes.

The experiments aim to analyze the classification accuracy when varying the percentage of genes selected by ReliefF and Gain ratio from 5% to 100% with a step size of 5%. For the purpose of this paper, the key question is how many genes should be selected to perform the best with microarray gene-expression data. Besides, we are interested in investigating whether or not the optimal percentage of genes depends on the characteristics of each classifier.

Note that the classification accuracy is just the number of samples being correctly classified, but this is not the most appropriate in the case of cancer classification problems. To discriminate between normal and cancerous data, it is especially important to take care of the false-positives and the false-negatives in order to perform a thorough comparison on the performance of different methods. False-positives are tolerable since further clinical experiments will be done to confirm the initial cancer diagnosis, but false-negatives are extremely detrimental because an ill patient might be misclassified as healthy.

## 4 Results and Discussion

Figure 1 shows the plot between accuracy rates and the percentage of the top-ranked genes for each database. It is found that all classifiers provide the highest accuracy using less than 20% of genes, irrespective of the feature selection algorithm. Examination of this figure reveals that in general, the RBF neural network and the naive Bayes classifier are the models most affected by the use of a large number of genes. For instance, in the Breast database the accuracy of RBF with the 5% top-ranked genes selected by ReliefF is 83.51%, but it significantly drops down to 52.58% when using the whole set of genes. Similarly, in the Colon



**Fig. 1.** Plots of the classification accuracy rates when varying the percentage of genes selected by the ReliefF (left) and Gain ratio (right) ranking algorithms

database the NBayes accuracy goes down from 77.42% with the 5% top-ranked genes selected by the Gain ratio to 58.06% with the total number of genes. It is also interesting to remark that the SVM has shown superior performance in

most cancer classification problems, probably because of its ability to deal with high-dimensional data and its robustness to noise [3,11], and also because all these data sets are linearly separable [2].

At this point, it could be especially interesting to show the relationship between the number of genes and the amount of samples in order to better understand how the ‘large  $G$ , small  $n$ ’ problem affects the classification results of gene-expression microarrays. To this end, the average number of samples per dimension (genes) for each database has been plotted in Fig. 2. This corresponds to the T2 data complexity measure [10], which describes the density of spatial distributions of samples by comparing the number of samples in the data set to the number of genes, ( $n/G$ ). As can be seen, there exists a negative correlation between the percentage of genes and the T2 measure, that is, higher values of  $X$  (% genes) are associated with lower values of  $Y$  (T2). This shows that, although the values of T2 are extremely small in all cases, the underlying difficulty of gene-expression microarray classification increases as the number of genes increases, which explains the decreasing tendency of accuracies presented in Fig. 1.

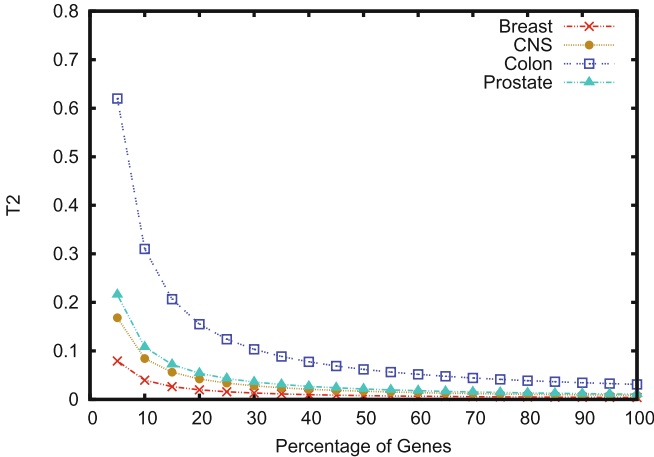


Fig. 2. Values of T2 when varying the percentage of genes

As already pointed out in Sect. 3, the false-negatives are even more relevant than the classification accuracy when assessing the performance of models for cancer classification based on gene-expression microarrays. Accordingly, Tables 2 and 3 report the false-negative rates given by each classifier both with the whole set of genes (100% of genes available) and the subset of genes that performed the best in terms of accuracy. The best result for each pair (database, classifier) is highlighted in bold. It is observed that the false-negative rate achieved with the best subset of genes is lower than that using 100% of genes in most cases: 22 out of 24 (4 data sets  $\times$  6 classifiers) with the ReliefF algorithm and 19 out of 24 with the Gain ratio feature ranking approach. These results corroborate

**Table 2.** False-negative rates with the ReliefF algorithm.

	1-NN		C4.5		SVM		RBF		NBayes		HAM	
	Best	100%	Best	100%	Best	100%	Best	100%	Best	100%	Best	100%
Breast	<b>0.196</b>	0.543	<b>0.283</b>	0.413	<b>0.217</b>	0.348	<b>0.109</b>	0.717	<b>0.196</b>	0.503	<b>0.283</b>	0.543
CNS	<b>0.077</b>	0.359	<b>0.179</b>	0.359	<b>0.077</b>	0.179	<b>0.179</b>	0.256	<b>0.231</b>	0.308	<b>0.179</b>	0.359
Colon	<b>0.318</b>	0.364	<b>0.182</b>	0.318	<b>0.227</b>	0.273	<b>0.182</b>	0.409	<b>0.182</b>	0.227	<b>0.227</b>	0.364
Prostate	<b>0.065</b>	0.130	<b>0.143</b>	0.156	<b>0.052</b>	0.078	0.273	<b>0.143</b>	<b>0.571</b>	0.675	0.256	<b>0.143</b>

**Table 3.** False-negative rates with the Gain ratio.

	1-NN		C4.5		SVM		RBF		NBayes		HAM	
	Best	100%	Best	100%	Best	100%	Best	100%	Best	100%	Best	100%
Breast	<b>0.217</b>	0.543	<b>0.348</b>	0.413	<b>0.239</b>	0.348	<b>0.413</b>	0.717	0.804	<b>0.503</b>	0.543	<b>0.503</b>
CNS	<b>0.282</b>	0.359	<b>0.154</b>	0.359	<b>0.128</b>	0.179	<b>0.231</b>	0.256	<b>0.256</b>	0.308	<b>0.256</b>	0.359
Colon	<b>0.227</b>	0.364	<b>0.273</b>	0.318	<b>0.227</b>	0.273	<b>0.182</b>	0.409	<b>0.136</b>	0.227	<b>0.273</b>	0.318
Prostate	<b>0.117</b>	0.130	<b>0.104</b>	0.156	<b>0.039</b>	0.078	0.299	<b>0.143</b>	0.675	0.675	0.503	<b>0.227</b>

the initial hypothesis that the removal of irrelevant (and redundant) genes leads to very significant gains in performance when the number of samples is large in comparison to the number of features, and it also produces a considerable decrease in computational requirements.

## 5 Concluding Remarks

The present paper has analyzed the effect of the high-dimensional, low-sample size problem for the classification of gene-expression microarrays. To this end, two feature ranking methods and six classifiers have been applied over four biomedical databases.

The experimental results have shown that the highest performance (as measured by the accuracy rate) was achieved by using a very small number of genes (in general, less than 20% of the total amount of genes), independently of both the gene ranking algorithm and the classifier. In addition, the T2 measure has shown that the complexity of classifying gene-expression microarrays increases as the amount of genes increases.

It has also been observed that RBF and naive Bayes appear to be the models most affected by (sensitive to) the ‘large  $G$ , small  $n$ ’ problem. On the other hand, the SVM with a linear kernel has performed the best in nearly all cases, probably because the experimental data sets are linearly separable. Finally, the false-negative rates have highlighted the benefits of using a subset with the top-ranked genes instead of the whole set because the presence of irrelevant genes may distort the classification problem in hand.

**Acknowledgment.** This work has partially been supported by the Spanish Ministry of Economy [TIN2013-46522-P], the Mexican PRODEP [DSA/103.5/15/7004], and the Generalitat Valenciana [PROMETEOII/2014/062].

## References

1. Alpaydin, E.: *Introduction to Machine Learning*. MIT Press, Cambridge (2010)
2. Bolón-Canedo, V., Morán-Fernández, L., Alonso-Betanzos, A.: An insight on complexity measures and classification in microarray data. In: *Proceedings of International Joint Conference on Neural Networks*, Killarney, Ireland, pp. 1–8 (2015)
3. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York (2000)
4. Dougherty, E.R.: Small sample issues for microarray-based classification. *Comp. Funct. Genomics* **2**(1), 28–34 (2001)
5. García, V., Sánchez, J.S.: Mapping microarray gene expression data into dissimilarity spaces for tumor classification. *Inform. Sci.* **294**, 362–375 (2015)
6. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439), 531–537 (1999)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
8. Heller, M.J.: DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.* **4**, 129–153 (2002)
9. Hira, Z.M., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* **2015**, 1–13 (2015). ID: 198363
10. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(3), 289–300 (2002)
11. Huang, L., Zhang, H.H., Zeng, Z.B., Bushel, P.R.: Improved sparse multi-class SVM and its application for gene selection in cancer classification. *Cancer Inform.* **12**, 143–153 (2013)
12. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., Raedt, L. (eds.) *ECML 1994*. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). doi:[10.1007/3-540-57868-4\\_57](https://doi.org/10.1007/3-540-57868-4_57)
13. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **9**(4), 1106–1119 (2012)
14. Lu, Y., Han, J.: Cancer classification using gene expression data. *Inf. Syst.* **28**(4), 243–268 (2003)
15. Raspe, E., Decraene, C., Berx, G.: Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise. *Semin. Cancer Biol.* **22**(3), 250–260 (2012)
16. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **53**(1–2), 23–69 (2003)
17. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**(19), 2507–2517 (2007)
18. Simon, R.: Analysis of DNA microarray expression data. *Best Pract. Res. Clin. Haematol.* **22**(2), 271–282 (2009)
19. Wang, L., Chu, F., Xie, W.: Accurate cancer classification using expressions of very few genes. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **4**(1), 40–53 (2007)
20. Zhang, C., Lu, X., Zhang, X.: Significance of gene ranking for classification of microarray samples. *IEEE-ACM Trans. Comput. Biol. Bioinform.* **3**(3), 312–320 (2006)