Pattern Analysis in DNA Microarray Data through PCA-Based Gene Selection

Ricardo Ocampo[1], Marco A. de Luna[1], Roberto Vega[1], Gildardo Sanchez-Ante[1], Luis E. Falcon-Morales[1], and Humberto Sossa[2]

[1] Tecnológico de Monterrey, Campus Guadalajara
Av. Gral Ramon Corona 2514
Zapopan, Jal, 45201, México
[2] Instituto Politécnico Nacional-CIC
Av. Juan de Dios Batiz S/N, Gustavo A. Madero 07738
México, Distrito Federal, México

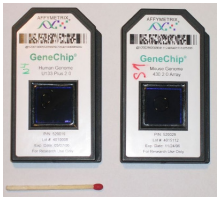
Abstract. DNA microarrays is a technology that can be used to diagnose cancer and other diseases. To automate the analysis of such data, pattern recognition and machine learning algorithms can be applied. However, the *curse of dimensionality* is unavoidable: very few samples to train, and many attributes in each sample. As the predictive accuracy of supervised classifiers decays with irrelevant and redundant features, the necessity of a dimensionality reduction process is essential. In this paper, we propose a new methodology that is based on the application of Principal Component Analysis and other statistical tools to gain insight in the identification of relevant genes. We run the approaches using two benchmark datasets: Leukemia and Lymphoma. The results show that it is possible to reduce considerably the number of genes while increasing the performance of well known classifiers.

1 Introduction

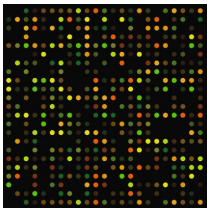
A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Each spot contains a clone of a gene. All genes in the microarray go through a process called hybridization which may change their color. Fluorescence measurements are made with a microscope. These measurements are used to determine the relative abundance of the sequence of each specific gene in the mRNA or DNA samples [2]. Microarray images typically contain several thousands of small spots, each of which represents a different gene in the experiment. Figure 1a shows one commercial chip, and Figure 1b shows an example of a microarray spot image with two different color dyes.

Although biopsy is still a standard diagnostic method for cancer, DNA microarrays are becoming an alternative. One of the main advantages of microarrays is the huge amount of molecular information that can be extracted and integrated to find common patterns within a group of samples. The two main goals of microarray studies are: 1) to identify molecular signatures associated with known classes, and 2) to discover new classes. Machine learning

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 532–539, 2014.



(a) Affymetrix chips



(b) Fluorescence spot image

Fig. 1. DNA Microarrays (images from Wikimedia Commons)

techniques can help via supervised methods (first goal), or unsupervised methods (second goal). Within the context of DNA Microarray, this basically means to identify discriminant genes. This is of fundamental interest in Biology and Medicine [9], [8].

A common problem in machine learning is *Feature Selection*, which consists in finding ways to reduce the dimensionality n of the feature space F, to reduce the risk of over-fitting as well as to allow efficient computation in the classifier. This is a main topic in general machine learning research since some time ago [14]. The approaches try, by different means, to identify and retain those attributes that better represent the original information contained in every sample of a dataset. In other words, the idea is to retain a subset F^* of F such that $||F^*|| << ||F||$ and that the elements of F^* still represent F reasonably well.

However, the processing of DNA microarray data, is particularly complex given that there are only a few tens of samples, and each sample contains several thousands of attributes. This is called the *curse of dimensionality*. Almost all of the datasets of DNA microarray data available up to now have such characteristic [1], [7], [12]. Reducing this dimensionality becomes crucial.

In this paper, we propose a methodology to reduce the number of attributes required to classify microarray data, based on Principal Component Analysis (PCA). We consider that this method will enable obtaining more insight in cancer characterization via DNA gene expression analysis. The rest of the paper is organized as follows: Section 2 describes an overview of different computational methods that have been used to perform feature selection as well as classification for DNA microarray analysis. Section 3 presents our methodology, Section 4 describes experiments and results and finally, Section 5 presents the conclusions and future work.

2 Previous Work

The methods for feature selection can be classified using different criteria [23]. They can be divided in univariate or multivariate methods, or, they can be separated in: filter and wrapper approaches. Filter methods usually use some measure to rank the attributes based on univariate functions, and then, the best ranked attributes are selected [15]. Wrapper methods are usually multivariate and they involve also a learning algorithm to evaluate the sets of attributes

used [13], [17]. Usually the first choice is to try a filter approach, since it is simple to run and requires $O(n)$ time. However, the main disadvantage is that it creates redundancy and evaluates attributes based on their individual scores, ignoring their relevance in combination with other attributes [18]. There are several examples on how this approach works on DNA microarrays data, like the one reported in [4] where the authors explore the use of PCA, Class-separability, Fisher ratio and t-test with a Support Vector Machine (SVM). Their results show that t-test allowed the classifier to perform a very good selection of genes. In [25] the authors test a three phase process for feature selection, using filters and Markovian tools. In [24], the authors perform a comparison of three alternatives using feature-ranking filters, correlation, and a wrapper.

There are also hybrid approaches, where a filter is applied and then a refinement process through a wrapper, like in [5], where a Minimum Redundancy-Maximum Relevance (MRMR) filter is used and then a Genetic Algorithm is applied to select the highly discriminant genes. In [19], a PAC (Probably Approximately Correct)-Bayes is proposed in combination with an SVM. Ant Colony Optimization [26], Genetic Algorithms and Neural Networks [22], Fuzzy-based [16], or as an optimization problem [6].

3 Methodology

The methodology proposed in this paper involves six steps, and it is based on using Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. However, although PCA compresses the information contained in a number p of original variables into a smaller set of q factors [10], each factor is a linear combination of all the p original variables. Therefore, PCA does not actually reduce the number of attributes, it only creates a different representation of the same data. Our methodology uses PCA as an intermediate step to reduce the number of attributes used in the analysis, and it can be summarized as follows:

1. Perform PCA and select the first q components.
2. Apply logistic regression using the q components and identify the d most relevant ones.
3. Analyze the d selected components as a group and eliminate the attributes whose weights are below a defined threshold.
4. Create a new dataset using only the remaining attributes of the last step.
5. Perform PCA on the new dataset and select the first q_2 components.
6. Classify the patterns.

3.1 Datasets

We used two publicly available, bi-class datasets in order to test our proposed methodology: The Leukemia Dataset [7] and the Lymphoma Dataset [20]. Each one contains 7,129 genes. The first dataset is divided in two subsets: training set

with 38 samples, and test set with 34 samples. The second dataset contains a total of 77 samples. We randomly divided it in a training set with 22 samples (11 of each class), and a test set with the remaining 55 samples.

3.2 Principal Component Analysis

The basic idea behind PCA is that, unless there is perfect correlation between two or more of the variables, p principal components are required to account for the p-dimensional variable space. PCA replaces the p original variables by a smaller number, q, of derived variables, the principal components, which are linear combinations of the original variables. Often, it is possible to retain most of the variability in the original variables with q much smaller than p. PCA projects p-dimensional data into a q-dimensional sub-space ($q \leq p$) in a way that minimizes the sum of squared distances from the points to their projections.

3.3 Logistic Regression

Logistic regression is a statistical technique used when the dependent variable is categorical. This technique is limited to bi-class problems, and focuses on identifying the independent variables that impact class membership in the dependent variable. Its basic model is described by:

$$Logit_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{1}$$

where $Logit_i$ represents the logit transformation used with the dependent variable of the sample i, X_n represents the nth attribute, and b_n represents its corresponding coefficient. The higher the absolute value of the coefficient, the higher the influence of the corresponding attribute for the class membership decision. An explanation of how to implement logistic regression is available in [10].

3.4 Classification

We implemented a Lattice Neural Network with Dendritic Processing (LNNDP) using the training method proposed by Sossa and Guevara in [21]. One of the advantages of this method is that it requires no parameter configuration. At the same time, it does not require random initialization values.

For the case of Support Vector Machines, we used the LIBSVM Library and trained the classifier as suggested in [3]. We used two different kernels: linear and radial basis function (RBF). In order to find the best parameters, we divided the training set in two parts: one for training, and the other for cross validation. The cross validation set was composed of 33% of the elements of the samples in the training set, chosen randomly.

For the case of Extreme Learning Machine [11], we used the basic ELM implementation of the Nanyang Technological University available at `http://www.ntu.edu.sg/home/egbhuang/elm_codes.html`. The only parameter to configure in this implementation is the number of neurons in the hidden

layer. We used a similar methodology than the one used for the SVM. We divided the train set in two subsets and then used cross validation to find the best number of neurons to use. To select the number of neurons we searched in the range [1,100].

4 Experiments and Results

We implemented PCA on the normalized training set and selected the q first components needed to retain at least 90% of the variance in order to create a new training set Q with the same amount of samples, but each being represented by q attributes (the chosen principal components). In the case of the Lymphoma dataset $q = 15$, while in the Leukemia dataset $q = 27$. Given that our interest is to identify which of the q components discriminate between the two classes of the dataset, we applied logistic regression over the training set Q and, retained the components whose coefficient's magnitude were above a certain treshold α. Both, α and the percentage of variance retained were chosen empirically. Table 1 shows in bold the coefficients selected in the Lymphoma dataset (components 3, 5, 6 and 10). A similar procedure was implemented over the Leukemia dataset. In this last case we selected components 3, 11, and 26.

Table 1. Coefficients obtained after applying logistic regression to the training set Q of the Lymphoma dataset

Coefficient	Magnitude	Coefficient	Magnitude	Coefficient	Magnitude
b_1	-0.0967	b_6	**0.4252**	b_{11}	0.2309
b_2	0.0452	b_7	0.2781	b_{12}	0.1347
b_3	**0.4591**	b_8	-0.0559	b_{13}	-0.0543
b_4	-0.1747	b_9	0.0275	b_{14}	0.03943
b_5	**0.3073**	b_{10}	**-0.4174**	b_{15}	0.1210

Each principal component has the form: $z_i = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$, where z_i is the ith principal component obtained using PCA, x_n is the nth attribute and θ_n is its corresponding weight. In order to reduce the number of attributes, we define a threshold t and apply the following rule:

$$\theta_i = \begin{cases} \theta_i, \texttt{if} \;\; |\theta_i| > t \\ 0, \;\; \texttt{if} \;\; |\theta_i| \leq t \end{cases}, i \in \{1, 2, \ldots, n\} \tag{2}$$

Thus, the threshold t was determined by analyzing the coefficients' magnitude distribution of the selected components based on a box-plot. We set $t_{lymphoma} = 0.006$ and $t_{leukemia} = 0.01$ and analyzed how many attributes (genes) had a coefficient different than zero in each component and analyzed which of them where present in more than one component. The results are shown in Table 2. By removing the genes with less contribution to the class discrimination capability of the logistic regression, we were able to reduce the dimensionality of the datasets from 7,129 attributes to only 972 for Lymphoma and 422 for Leukemia, which represent 13.63% and 5.91% of the original ones. In addition, these attributes

Table 2. Number of genes that are present in different components

Dataset	Threshold	0-times	1-time	2-times	3-times	4-times
Lymphoma	0.006	236	1117	2296	2508	972
Leukemia	0.01	1846	2861	2000	422	-

Table 3. Comparison of classification precision using PCA over the original Lymphoma(above) and Leukemia(below) datasets, and using PCA over the proposed reduced dataset. In bold the highest precision of each method.

q_2	LNNDP		SVM RBF		SVM LIN		ELM	
	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.
3	74.55%	65.46%	60.49%	50.82%	63.69%	53.64%	64.36%	57.40%
5	69.09%	**83.64%**	64.00%	**65.15%**	70.76%	69.06%	62.47%	66.13%
7	60.00%	81.82%	64.91%	63.67%	72.47%	**73.95%**	64.26%	66.64%
15	61.82%	72.73%	49.75%	59.95%	71.31%	71.22%	67.38%	**68.22%**

q_2	LNNDP		SVM RBF		SVM LIN		ELM	
	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.
3	79.41%	76.47%	69.73%	80.15%	70.56%	**85.88%**	66.73%	**79.27%**
5	73.53%	76.47%	71.00%	77.29%	71.15%	78.50%	69.09%	75.47%
7	70.59%	76.47%	72.09%	77.97%	71.74%	79.62%	66.68%	76.77%
15	73.53%	**88.24%**	77.35%	**82.59%**	78.21%	83.24%	72.68%	79.21%

might have biological and medical significance, considering that they represent the actual genes. We then created a new Lymphoma dataset, and a new Leukemia dataset, using only the genes present in all the selected components.

Having reduced the actual number of attributes in the dataset, we implemented again PCA over the new datasets and tested different classification algorithms on the test set using the first 3, 5, 7 and 15 components, that is, q_2. For comparison purposes, we also implemented the classification algorithms over the PCA implemented over the original dataset with 7,129 attributes using the same number of components. The results are shown in Table 3. These results suggest that classifying the patterns directly after the implementation of PCA yields suboptimal results. By reducing the number of genes used we not only reduced the computational cost of the analysis, but also improved the accuracy of all the tested classification methods. In some cases the improvement was above 10%. LNNDP is a relatively new method for classification that outperforms the other well known methods.

5 Conclusions

PCA is a common tool in pattern recognition used to reduce the dimensionality of a dataset. However, it uses all the original attributes to create a reduced set of factors. Therefore, we have the same number of attributes, but represented

in a different way. Besides, because PCA uses no information of the class labels, there is no guarantee that the variance retained by the first k components is the variance needed to discriminate among classes. Using the proposed methodology we related the principal components to the desired class labels and determined which attributes have more influence to perform discrimination. We diminished the number of attributes from 7,129 to less than a thousand attributes in both datasets while improving the precision performance of the classification algorithms by 15% in the best case. One remarkable point of this methodology is that it found that, at least in the dataset used for the experiments, the most relevant components were not the first ones. This finding suggest that applying PCA over a dataset and use the first k components is not enough to achieve optimal classification; however, using it as a tool for attribute reduction using the methodology here proposed could improve the performance of the classifiers.

Acknowledgments. The authors thank Tecnológico de Monterrey, Campus Guadalajara, for their support under the Research Chair in Information Technologies and Electronics, as well as IPN-CIC under project SIP 2014-0776, and CONACYT under project 155014 for the economical support to carry out this research.

References

1. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769), 503–511 (2000)
2. Brown, P.O., Botstein, D.: Exploring the new world of the genome with DNA microarrays. Nature Genetics 21, 33–37 (1999)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
4. Chu, F., Wang, L.: Applications of support vector machines to cancer classification with microarray data. Int. Journal of Neural Systems 15(06), 475–484 (2005)
5. Akadi, A.E., Amine, A., El Ouardighi, A., Aboutajdine, D.: A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems 26(3), 487–500 (2011)
6. Gardeux, V., Natowicz, R., Wanderley, M.F.B., Chelouah, R.: Optimization for feature selection in DNA microarrays. Heuristics: Theory and Applications (2013)
7. 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.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)
8. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182 (2003)
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning 46(1-3), 389–422 (2002)
10. Hair, J., Black, W., Babin, B., Anderson, R.: Multivariate data analysis, 7th edn. Prentice Hall, USA (2010)

11. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489–501 (2006)
12. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7(6), 673–679 (2001)
13. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1), 273–324 (1997)
14. Koller, D., Sahami, M.: Toward optimal feature selection. Tech. rep., Stanford InfoLab, Stanford University (1996)
15. Liu, H., Setiono, R.: A probabilistic approach to feature selection-a filter solution. In: ICML, vol. 96, pp. 319–327. Citeseer (1996)
16. Park, D., Jung, E.-Y., Lee, S.-H., Lim, J.: A composite gene selection for dna microarray data analysis. Multimedia Tools and Applications, 1–11 (2013)
17. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S.: Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognition 39(12), 2383–2392 (2006)
18. Ryu, J., Cho, S.-B.: Towards optimal feature and classifier for gene expression classification of cancer. In: Pal, N.R., Sugeno, M. (eds.) AFSS 2002. LNCS (LNAI), vol. 2275, pp. 310–317. Springer, Heidelberg (2002)
19. Shah, M., Marchand, M., Corbeil, J.: Feature selection with conjunctions of decision stumps and learning from microarray data. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(1), 174–186 (2012)
20. Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., et al.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8(1), 68–74 (2002)
21. Sossa, H., Guevara, E.: Efficient training for dendrite morphological neural networks. Neurocomput. 131, 132–142 (2014)
22. Tong, D.L., Schierz, A.C.: Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. Artificial Intelligence in Medicine 53(1), 47–56 (2011)
23. Tsamardinos, I., Aliferis, C.F.: Towards principled feature selection: Relevancy, filters and wrappers. In: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (2003)
24. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W.: Gene selection from microarray data for cancer classification-a machine learning approach. Computational Biology and Chemistry 29(1), 37–46 (2005)
25. Xing, E.P., Jordan, M.I., Karp, R.M., et al.: Feature selection for high-dimensional genomic microarray data. In: ICML, vol. 1, pp. 601–608. Citeseer (2001)
26. Yu, H., Gu, G., Liu, H., Shen, J., Zhao, J.: A modified ant colony optimization algorithm for tumor marker gene selection. Genomics, Proteomics & Bioinformatics 7(4), 200–208 (2009)