Hybrid (Generalization-Correlation) method for feature selection in high dimensional DNA microarray prediction problems

Yasel Couce[1], Leonardo Franco[2], Daniel Urda[2], José L. Subirats[2], and José M. Jerez[2]

[1] Universidad de Ciencias Informáticas, La Habana, Cuba,
yaselc@uci.cu
[2] Universidad de Málaga, Department of Computer Science, ETSI Informática, Spain.
{lfranco, jlsubirats, durda, jja} @lcc.uma.es

Abstract. Microarray data analysis is attracting increasing attention in computer science because of the many applications of machine learning methods in prediction problems. The process typically involves a feature selection step, important in order to increase the accuracy and speed of the classifiers. This work analyzes the characteristics of the features selected by two wrapper methods, the first one based on artificial neural networks (ANN) and the second in a novel constructive neural network (CNN) algorithm, to later propose a hybrid model that combines the advantages of wrapper and filter methods. The results obtained in terms of the computational costs involved and the prediction accuracy reached show the feasibility of the hybrid model proposed here and indicate an interesting research line for the near future.

Keywords: DNA Microarray, Feature Selection, Data Mining, Constructive Neural Networks.

1 Introduction

DNA microarray technology has opened up new research directions and significant opportunities in biomedical sciences. DNA microarray technology makes possible to measure simultaneously the expressions levels of thousands of genes in a single experiment, providing unique and useful data for a wide range of experimental research, e.g., predicting disease outcome in patients. 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 data, microarray data analysis face the "large-p-small-n" paradigm [9] also known as the curse of dimensionality. Generally, most of these features are irrelevant to a specific study and represent noise for most of the prediction systems, therefore the application of machine learning techniques [5] are becoming increasingly needed in order to enhance the speed and accuracy in prediction systems.

The microarray data analysis usually involves a preprocessing step, which consists in the selection of features (genes) relevant for the classification step.

These feature selection algorithms are grouped into two categories, filter methods and wrapper methods [3]. Filter methods selects relevant features based on general characteristics of the training data. Wrapper methods requires an specific classifier algorithm to evaluate the suitability of each subset of features found. As a result of using the classifier in the feature subset evaluation, wrapper methods tend to outperform the prediction accuracy of the filter methods, but the constantly training of the classifier leads to a high computational cost for the wrapper methods, which makes them not much used in microarray data analysis [2]. On the other hand, filter methods are fast and computationally simple, making them more suitable for tasks on high-dimensional datasets.

A recent proposal made by Urda et al. [8] shows how constructive neural networks (CNN) algorithm can reduce time and increase accuracy in microarray data analysis, in particular in comparison to ANN. Using a novel constructive algorithm (C-Mantec) [7] to predict estrogen receptor status, Urda et al. [8] improves the results achieved by Lancashire et al. [4] using a stepwise forward selection artificial neural network approach. Despite the results obtained by [8], the solution still suffers the drawbacks of the wrapper methods making the search in feature space very time-consuming.

The combination of filter and wrapper approaches has led to alternative proposals [6][1] looking to overcome the disadvantages of the two methods separately. Hybrid models incorporate the relationship of the wrappers with the classifiers, in order to increase the accuracy prediction of the selected subset and use the analysis of the properties of the data set, performed by the filters, to achieve speed and scalability. In this work we present an hybrid wrapper-filter model using a constructive neural network algorithm to build a classifier using the information from the training patterns in order to facilitate its adaptation to a given problem. The proposal is based on the use of a recently introduced constructive neural network algorithm (C-Mantec) [7] and on the use of a simple correlation measure to rank the features in the dataset, in order to avoid redundancy in the selected features.

2 Materials and Methods

2.1 Materials

The dataset used in this work comes from the study published by West et al. [10] (http://data.cgt.duke.edu/west.php). This study used microarray technology to analyze primary breast tumors in relation to estrogen receptor (ER) status. The dataset contains the expression levels of a total of 7129 genes measured in 49 breast tumor samples (25 ER+ and 24 ER- cases).

2.2 Methods

We first analyze in this work two previous studies [4] [8] where feature selection is implemented and characterize the process measuring the prediction ability of

using the individual genes and the correlation between the selected features, to use later these two characteristics of the dataset to test three different strategies. For measuring the generalization ability of the individual genes for the selection process and for estimating the predictive accuracy of using all the selected variables we used C-Mantec, a constructive neural network algorithm that generates very compact neural architectures with state-of-the-art generalization capabilities [7] .

The well known linear correlation coefficient (r), computed for a pair of variables $(X;Y)$ and shown below in equation 1, was used for estimating the redundancy among a set of variables.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x}$ is the mean of X, and $\bar{y}$ is the mean of Y. The value of r belongs to interval $[-1, 1]$. If r is zero then X and Y are totally independent. The closer is the value of r to the extremes of the interval $[-1, 1]$, the closer X and Y to a perfect linear relationship. As we need to measure the redundancy of one feature to a set of features, we computed the correlation of variable X and a set of n variables $(Y_1, Y_2, \ldots, Y_n)$ as the mean of the correlation value of each pair (X, Y_i).

3 Experiments and Results

We have first analyzed the set of features selected in the studies by Lancashire et al. [4] and Urda et al. [8], where the datasets have been selected using forward selection methods based only in measuring prediction accuracy, in order to measure the relevance of individual generalization and correlation. Table 1 shows the features selected by Lancashire et al. with their respective measures of correlation and generalization, while table 2 shows the same analysis for the variables selected by Urda et al. Both tables show first the values for the correlation measured as an average between pairs of variables, indicating the obtained correlation coefficient, then a rank position between 1 and 7129 (number of features of the dataset) and a normalized rank between 0 and 1. The three rightmost columns of the table show the generalization ability obtained when the C-Mantec algorithm is trained only with a single variable, and the columns show the absolute value, rank position and normalized rank.

The data shown in the tables indicate that the variables selected in the Lancashire analysis present relatively high individual generalization values (average normalized rank equals to 0.1579) while correlation among variables is high (average normalized rank equals to 0.7614), indicating a large redundancy between the selected variables. For the features selected in the Urda et al. study the correlation between variables is lower (average normalized rank equals to 0.2247) but the generalization measure of the individual variables is higher (0.2990 for the normalized rank). Surprisingly, even if both sets of variables were chosen using a forward selection method the characteristics of the selected variables are

Table 1. Features selected in the work of Lancashire et al. [4] ranked according to the correlation measure used in the present work (see the text for more details).

		Correlation measure			Generalization measure		
	Probe Set ID	correlation coefficient	rank position	rank [0-1]	mean %	rank position	rank [0-1]
1	X58072_at	–	–	–	86.0	4	0.0004
2	Z29083_at	0.3302	6402	0.898	83.4	9	0.0011
3	M81758_at	0.1763	5033	0.706	52.6	4550	0.6382
4	M60748_at	0.1520	3515	0.493	60.2	1419	0.1989
5	M74093_at	0.1835	4590	0.644	81.8	15	0.0020
6	U22029_f_at	0.2175	5484	0.769	58.0	2242	0.3144
7	U96131_at	0.2657	6625	0.929	73.2	131	0.0182
8	M96982_at	0.2391	6353	0.891	64.4	642	0.0899
	mean	0.2235	5429	0.761	69.9	1126	0.1579

Table 2. Features selected in the work of Urda et al. [8] ranked according to the correlation coeffficient averaged among variables and the generalization ability (see the text for more details).

		Correlation measure			Generalization measure		
	Probe Set ID	correlation coefficient	rank position	rank [0-1]	mean %	rank position	rank [0-1]
1	X76180_at	–	–	–	85.6	6	0.0007
2	HG4749-HT5197_at	0.0566	1785	0.250	59.2	1742	0.2442
3	M31520_rna1_at	0.1043	1431	0.201	55.2	3383	0.4745
4	U20325_at	0.1265	1592	0.223	55.2	3399	0.4767
	mean	0.0958	1603	0.225	63.8	2132	0.2990

different. It is worth noting that the Lancashire et al. work used ANN as classification algorithm while in the Urda et. al. analysis C-Mantec algorithm was used. Given this discrepancy between the observed characteristics of the datasets, we have decided to propose and test **three** different strategies in order to develop a selection method:

Relevance only (ROnly) It is the simplest of the three strategies, aimed to select those features with the highest generalization over the test data subset. The pseudocode of this strategy is presented below as Algorithm 1.

Algorithm 1 Pseudocode of the ROnly algorithm.

for each feature f_i in dataset D **do**
    {Create model for f_i and compute its generalization performance on D}
    $g(i) \leftarrow$ **CMantec**$(f_i, T_0, I_{max}, g_{fac})$;
end for
$n \leftarrow$ **size**(D)
for $j = 1$ **to** 10 **do**
    {Select the feature with the highest generalization performance on D}
    $Set(j) \leftarrow \{f_i \in D : \forall x \in \{1, \ldots, n\}(g(i) \geqslant g(x))\}$;
    $g(i) \leftarrow 0$
end for
return Set

Relevance first, then redundancy (RelevanceF) This strategy consists in selecting the ten percent of the most relevant features from the data set (those with the highest generalization value over the test data subset) and then within that ten percent select the feature that is less redundant with the subset of variables already selected for classification. Algorithm 2 specifies the pseudocode for this strategy.

Algorithm 2 Pseudocode of the RelevanceF algorithm.

```
for each feature f_i in dataset D do
    {Create model for f_i and compute its generalization performance on D}
    g(i) ← CMantec(f_i, T_0, I_max, g_fac);
end for
n ← size(D)
{Select the ten percent of features with the highest generalization performance on D}
for j = 1 to n ÷ 10 do
    T(j) ← {f_i ∈ D : ∀x ∈ {1, ..., n}, (g(i) ⩾ g(x))};
    g(i) ← 0
end for
Set(1) ← T(1)
for j = 2 to 10 do
    {Select the feature less redundant to features in Set}
    Set(j) ← {f_i ∈ T : ∀x ∈ {1, ..., n}, (r̄(f_i, Set) ⩽ r̄(f_x, Set))};
end for
return Set
```

Redundancy first, then relevance (RedundancyF) Contrary to the previous strategy, this algorithm select first a ten percent of the features that are less redundant to the features already selected for classification and within this ten percent the most relevant attribute (high generalization) is chosen. Algorithm 3 shows pseudocode for this strategy.

Algorithm 3 Pseudocode of the RedundancyF algorithm.

```
for each feature f_i in dataset D do
    {Create model for f_i and compute its generalization performance on D}
    g(i) ← CMantec(f_i, T_0, I_max, g_fac);
end for
n ← size(D)
{Select the feature with the highest generalization performance on D}
Set(1) ← {f_i ∈ D : ∀x ∈ {1, ..., n}, (g(i) ⩾ g(x))};
g(i) ← 0
for j = 2 to 10 do
    {Select the ten percent of features less redundant to features in Set}
    for t = 1 to n ÷ 10 do
        R(t) ← {f_i ∈ D : ∀x ∈ {1, ..., n}, (r̄(f_i, Set) ⩽ r̄(f_x, Set))};
    end for
    {Select the feature with the highest generalization performance on D}
    Set(j) ← {f_i ∈ R : ∀x ∈ {1, ..., n}, (g(i) ⩾ g(x))};
end for
return Set
```

In all cases the C-Mantec algorithm was the classification method used, with the following values for the parameters: $g_{fac} = 0.2$ (network growing factor), $I_{max} = 100,000$ (maximum number of iterations), while T_0 (initial temperature) was set equal to the number of input variables. A ten-fold cross validation approach was used and the results obtained are indicated in table 3 and in figure 1. Table 3 shows the mean and the standard deviation across the set of ten observed values for the three algorithms, with best value highlighted. We can observe that the third algorithm (Redundancy first, then relevance) outperformed the other two achieving the highest generalization and also showing the

lowest standard deviation. (Figure 1), and thus we choose this algorithm as our proposal for the feature selection process.

Table 3. Mean and standard deviation observed for the three algorithms

Algorithm	Generalization C-Mantec
ROnly	89.18 ± 1.03
RelevanceF	84.72 ± 1.92
RedundancyF	92.82 ± 0.51

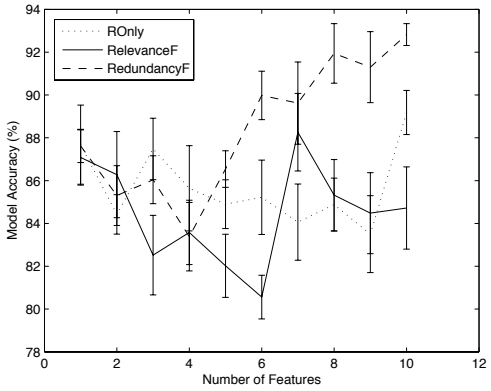

Fig. 1. Results of ten-fold cross validation of the C-Mantec algorithm applied to the test dataset using the features selected by the three algorithms. Error bars represent standard error of the mean expressed in percentage.

Table 4 shows the characteristic of the selected variables chosen by the RedundacyF algorithm. The average normalized rank is much lower for both correlation between features (0.069) and individual prediction accuracy (0.025) than in the previous analyzed studies [4] and [8].

As a final comparison, table 5 shows the generalization ability obtained using C-Mantec on a 10-fold cross validation scheme including all the selected features chosen by the three studies (Lancashire et al., Urda et al., and the present proposal) . Given that the proposed approach (RedundancyF) only uses C-Mantec in the first phase of the algorithm (estimation phase) and that the second phase (selection phase) consumes an average of 0.4 seconds of CPU time to find each new feature, the execution time remains almost invariable as the number of selected features increases, while this is not the case for the previous proposals [8] and [4]. This make our algorithm a more feasible choice to scale on high-dimensional datasets. A simple mathematical analysis shows below that for the forward selection methods used in [4] [8] the computational costs approximately scale with both the total number of features available (N_I) and with the number

Table 4. Features selected by RedundacyF algorithm, ranked according to the correlation measure used in the present work (see the text for more details).

		Correlation measure			Generalization measure		
	Probe Set ID	correlation coefficient	rank position	rank [0-1]	mean %	rank position	rank [0-1]
1	X03635_at	–	–	–	87.6	1	0
2	M85289_at	0.0052	234	0.033	72.0	157	0.0219
3	U60269_cds2_at	0.0250	371	0.052	67.4	379	0.0530
4	L08044_s_at	0.0857	540	0.076	79.4	32	0.0043
5	U41371_at	0.0761	354	0.050	69.4	254	0.0355
6	L13278_at	0.0884	446	0.062	68.2	317	0.0443
7	X06614_at	0.1137	563	0.079	73.2	132	0.0184
8	L37199_at	0.1147	651	0.091	72.0	156	0.0217
9	HG2279-HT2375_at	0.1186	631	0.088	73.8	116	0.0161
10	S77410_at	0.1184	649	0.091	69.4	251	0.0351
	mean	0.0829	493	0.069	73.2	179	0.0250

Table 5. Percentage accuracy using C-Mantec on a 10-fold cross validation scheme using all the selected features chosen in each work.

Method	Percentage accuracy
Urda et al. [8]	0.950 ± 0.103
Lancashire et al. [4]	0.946 ± 0.091
RedundancyF	0.922 ± 0.099

of features to be selected (N_V), while for the hybrid method proposed the computationally costs scale linearly only with the total number of available features. The following equations shows the calculation involved:

$$CPU_{time}(FSel) = \sum_{i=1}^{V}(N_I - i + 1)N_V T_{gen}(i) \sim N_I N_V \overline{T_{gen}} \quad (2)$$

$$CPU_{time}(Hybrid) = N_I(T_{gen}(V = 1) + N_V T_{cor}) \sim N_I T_{gen}(V = 1), \quad (3)$$

where $T_{gen}(i)$ is the time needed to compute the generalization ability of a given model using i input variables ($\overline{T_{gen}}$ indicates averaging). For the hybrid model T_{cor} indicates the CPU time needed to compute a correlation measure between pair of variables. In our case T_{cor} was very small and thus the total CPU time depends mainly on the computational cost of computing the generalization of the model. The validity of the previous analysis was checked for some particular values but lack of space in the present work leaves a more detailed analysis to be published elsewhere.

4 Conclusions and Further Work

In this work we have first carried out an analysis of the characteristics of the features selected by two recent publications, to further propose and test three strategies for selection of informative genes in DNA microarray experiments through applying a hybrid model of constructive neural network algorithm and a simple correlation-based algorithm. Even though the new introduced models do not reach the level of effectiveness of the reviewed approaches (Table 5), we must emphasize that a major goal of this research was firstly to reduce the computational cost of the feature selection task and on the other hand to analyze whether both correlation between features and level of generalization of

individual features are important characteristics for the feature selection task, fact that the obtained results seems to confirm. Further work will be centered on extensions of the RedundancyF algorithm in order to increase its speed and prediction accuracy on DNA microarray prediction problems. Better measures for estimating feature redundancy will be tested, as they may permit to capture nonlinear correlation effects, and additionally, tests on different databases using different classifiers will be conducted to fully validate the approach.

Acknowledgements

The authors acknowledge support from CICYT (Spain) through grants TIN2008-04985 and TIN2010-16556 (including FEDER funds) and from Junta de Andalucía through grant P08-TIC-04026.

References

1. Huda, S., Yearwood, J., Strainieri, A.: Hybrid Wrapper-Filter Approaches for Input Feature Selection Using Maximum Relevance and Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA). In: 4th International Conference on Network and System Security. pp. 442 –449 (2010)
2. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine 31, 91 – 103 (2004)
3. Kohavi, R., John, G.: The Wrapper approach. In: Feature Extraction, Construction and Selection: a data mining perspective pp. 33–51 (1998)
4. Lancashire, L.J., Rees, R.C., Ball, G.R.: Identification of gene transcript signatures predictive for er and lymph node status using a stepwise forward selection ann modelling approach. Artif. Intell. Med. 43, 99–111 (2008)
5. Pirooznia, M., Yang, J., Yang, M.Q., Deng, Y.: A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 9, S13 (2008)
6. Sebban, M., Nock, R.: A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recognition 35, 835 – 846 (2002)
7. Subirats, J.L., Jerez, J.M., Gómez, I., Franco, L.: Multiclass Pattern Recognition Extension for the New C-Mantec Constructive Neural Network Algorithm. Cognitive Computation 2, 285–290 (2010)
8. Urda, D., Subirats, J., Franco, L., Jerez, J.: Constructive neural networks to predict breast cancer outcome by using gene expression profiles. In: Lecture Notes in Artificial Intelligence, 6096. Proceedings of the IEA-AIE. pp. 317–326 (2010)
9. West, M.: Bayesian factor regression models in the "large p, small n" paradigm. Bayesian statistics 7, 723–732 (2003)
10. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 98, 11462–11467 (2001)