Progress in Artificial Intelligence (2019) 8:63–71
https://doi.org/10.1007/s13748-018-0148-6

REGULAR PAPER

# Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory

Laura Cleofas-Sánchez[1] · J. Salvador Sánchez[2] · Vicente García[3]

Received: 22 January 2018 / Accepted: 22 March 2018 / Published online: 31 March 2018


**Abstract**
In general, gene expression microarrays consist of a vast number of genes and very few samples, which represents a critical challenge for disease prediction and diagnosis. This paper develops a two-stage algorithm that integrates feature selection and prediction by extending a type of hetero-associative neural networks. In the first level, the algorithm generates the associative memory, whereas the second level picks the most relevant genes. With the purpose of illustrating the applicability and efficiency of the method proposed here, we use four different gene expression microarray databases and compare their classification performance against that of other renowned classifiers built on the whole (original) feature (gene) space. The experimental results show that the two-stage hetero-associative memory is quite competitive with standard classification models regarding the overall accuracy, sensitivity and specificity. In addition, it also produces a significant decrease in computational efforts and an increase in the biological interpretability of microarrays because worthless (irrelevant and/or redundant) genes are discarded.

**Keywords** Associative memory · Gene selection · Disease prediction · Gene expression microarray

## 1 Introduction

Gene expression microarray is a high-throughput genomic technology in research and clinical management that allows to record and monitor the expression levels of thousands of genes simultaneously within a few different samples. The expression level of a gene can be viewed as an estimate of the concentration of its mRNA transcript in a cell at a given time. The primary objective of using microarrays is to classify or predict the category of a sample based on its gene expression profile. A plethora of computational methods have been applied to the analysis of microarrays [18,31], and in particular to discriminate between cancerous and non-cancerous tissues, to characterize distinct types or subtypes of tumors and even to predict the reaction to a specific therapeutic drug and the risk of relapse [9,33,37,46].

Nonetheless, the use of microarrays for classification poses a crucial computational challenge arising from the huge number of genes ($G$) and the limited quantity of samples ($n$) [14]. The number of genes is usually of the order of thousands, but the number of samples is below a hundred. This problem is known as the 'large $G$, small $n$' or 'curse of dimensionality' phenomenon, which increases the difficulty of classification significantly, degrades the generalization capability of classifiers and hinders the understanding of the relationships between genes and samples [13,16,23,38]. Moreover, only a few relevant genes are needed [5,7,17,47]. The most common practice to address this question is to use some feature (gene) selection method by choosing a small portion of informative variables that contribute most to any subsequent task for clinical identification, classification, prediction or interpretation.

Feature selection focuses on the deletion of irrelevant, noisy and redundant genes from the microarrays with the

✉ J. Salvador Sánchez
sanchez@uji.es

Laura Cleofas-Sánchez
laura18cs77@gmail.com

Vicente García
vicente.jimenez@uacj.mx

[1] National Institute of Genomic Medicine, 14610 Ciudad de México, Mexico

[2] Department of Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain

[3] Multidisciplinary University Division, Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, Chihuahua, Mexico

aim of preserving the genes that best discriminate samples of different classes (tissue categories, disease states or clinical outcomes). Besides, it also helps to identify patterns of gene expression associated with a particular disease. On the other hand, it has to be remarked that the need of applying feature selection is imperative in the case of gene expression microarrays not only owing to the extremely large number of input variables, but also because a considerable amount of them can be highly correlated with other variables. Over the last decades, many different algorithms have been developed for gene selection using filter, wrapper, embedded and hybrid methods [5,7,19,30,35,39,43,49].

In the present paper, we introduce a technique to classify the gene expression microarray data with a two-stage associative memory. The first stage involves the construction of a hybrid associative neural network, whereas the second stage allows for the selection of the most relevant (differentially expressed) genes for the classification of the tissue samples. Thus, the purpose of this study is twofold: (i) to analyze the practicability of this hybrid connectionist model when applied to the classification of gene expression microarrays and (ii) to compare its performance with that of several standard prediction algorithms that are widely used in the biomedical domain.

Following the general definition given in the literature, an associative memory constitutes a content-addressable neural network based upon matrix algebra [27] that connects each input vector $\mathbf{x}$ with its corresponding output vector $\mathbf{y}$. In practice, the associative memory takes the form of a connection weight matrix $\mathbf{W} = [\mathbf{w_{i,j}}]_{\mathbf{C}\times\mathbf{G}}$ built from a finite set of $n$ encoded associations, generally known as fundamental set of associations, $\{(\mathbf{x}^\mu, \mathbf{y}^\mu) \mid \mu = 1, 2, \ldots, n\}$, where $\mathbf{x}^\mu \in \mathbb{R}^G$ correspond to the fundamental input samples of dimension $G$ and $\mathbf{y}^\mu \in \mathbb{R}^C$ are the fundamental output samples of dimension $C$. Then $x_j^\mu$ represents the $j$-th element of an input sample $\mathbf{x}^\mu$, and $y_i^\mu$ indicates the $i$-th element of an output sample $\mathbf{y}^\mu$.

In general, an associative memory can be of two types depending on the nature of memorized associations: hetero-associative (e.g., lernmatrix [41], linear associator [4,26], alpha-beta associative memory [48]) and auto-associative (e.g., associatron [36], Hopfield network [21], chaotic neural network [2], bidirectional associative memory [28]). The hetero-associative memories connect input samples with output samples of different nature and formats (i.e., $\mathbf{x}^\mu \neq \mathbf{y}^\mu$), whereas the auto-associative neural networks are viewed as a particular case of the former where $\mathbf{x}^\mu = \mathbf{y}^\mu$ and $G = C$. On the other hand, while the auto-associative models include a single layer in which all processing units are completely connected by feedback links, the hetero-associative memories consist of more than one layer and each layer is fully connected to all the others.

Since the seminal works of Steinbuch [41], Anderson [4], Kohonen [26] and Nakano [36], associative memories have gradually been the subject of many theoretical and empirical studies. For instance, the bidirectional associative memory was applied to the diagnosis of cancer based on the elemental contents in serum samples [50]. Chartier and Lepage [11] developed a modified Hopfield network to learn and detect edges from gray level images. Arya et al. [6] implemented a face recognition system using auto-associative memory blocks in parallel. A technique for segmentation and classification of soft tissues from textural features of medical images based on a bidirectional associative memory was designed by Sharma et al. [40]. Sudo et al. [42] introduced a new associative memory capable of realizing both bidirectional and multidirectional associations. Aldape-Pérez et al. [3] proposed the use of an associative neural network for medical decision support systems. Aghajari et al. [1] suggested a chaotic hetero-associative memory built using a learning strategy that allows to store and recall a set of associated patterns even when these are noisy. Vaishnavi et al. [44] adapted the Hopfield network for isolated word recognition. Villuendas-Rey et al. [45] presented the naïve associative classifier, which was based upon a new similarity operator with the capability to handle missing values and both numerical and categorical data. Cleofas-Sánchez et al. [12] built an associative memory with a translation of the coordinate axes for financial distress prediction, showing higher performance than other classifiers, especially when the data sets exhibited some overlapping and severe imbalance in class distribution.

Henceforth, the article is organized as follows. Section 2 provides a detailed description of the two-stage hetero-associative neural network that we propose here. The databases and experimental setup are outlined in Sect. 3, whereas the classification results are reported and discussed in Sect. 4. Finally, Sect. 5 summarizes the most remarkable findings that can be gathered from this study and identifies some avenues for further research.

## 2 The two-stage associative memory

The two-level algorithm presented in this paper is based on a particular implementation of an hetero-associative memory, which combines the learning stage of the linear associator with the recall stage of the Steinbuch's lernmatrix. This approach differs from the hybrid associative memory introduced by Cleofas-Sánchez et al. [12] in that the former incorporates a feature selection stage into the original model, whereas this was merely designed for class prediction. Since our method merges feature selection and classification based on associative memories, it will be here referred to as hybrid hetero-associative memory (HAM).

It is worth highlighting that the main advantages of this model over the linear associator and the lernmatrix are twofold: (i) it can work with real-valued input samples, whereas the lernmatrix only accepts the values 0 and 1; and (ii) the input vectors are not constrained to be orthonormal, unlike the linear associator. Besides, it has to be mentioned that the hetero-associative memories are generally considered to be tolerant to noise on the input and incomplete stimuli.

### 2.1 Level 1: Construction of the hetero-associative memory

The first level of the HAM algorithm is devoted to building a matrix $\mathbf{W}$ such that when an input sample $\mathbf{x}^\mu$ is presented, the stored sample $\mathbf{y}^\mu$ associated with it will be retrieved. This matrix construction process consists of two sequential steps:

1. For each encoded association $(\mathbf{x}^\mu, \mathbf{y}^\mu)$, calculate the outer product $\mathbf{y}^\mu(\mathbf{x}^\mu)^T$, where $(\mathbf{x}^\mu)^T$ corresponds to the transpose of $\mathbf{x}^\mu$.
2. Sum the $n$ outer products to yield the matrix $\mathbf{W} = \alpha \sum_{\mu=1}^{n} \mathbf{y}^\mu(\mathbf{x}^\mu)^T$, where $\alpha$ denotes the normalizing constant (usually set to $1/n$). The $(i, j)$-th element of $\mathbf{W}$ is given by

$$w_{i,j} = \sum_{\mu=1}^{n} y_i^\mu x_j^\mu.$$

Moreover, it has to be pointed out that the learning stage of this hetero-associative memory starts by translating the coordinate axes to an origin located at the centroid of the fundamental input samples. The purpose of moving the fundamental set is to represent the input samples in a new $G$-dimensional space where samples that belong to two different classes are located diametrically opposite to each other, and the midpoint of the diameter is given by the mean vector $\overline{\mathbf{x}}$, that is, the origin of the new coordinate axes. This translation should provide higher classification performance because it is expected that samples of different classes will probably be put quite far apart in separate quadrants.

Let $A = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ be a finite set of $n$ fundamental input samples, let $C$ denote the number of classes, and let $\hat{A} = \{\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots, \hat{\mathbf{x}}^n\}$ be the set of fundamental input samples translated to the new space. Then the pseudo-code of the learning stage to build the connection weight matrix $\mathbf{W}$ is presented in Algorithm 1, which is based on the associative memory proposed by Cleofas-Sánchez et al. [12].

**Algorithm 1** Construction stage

```
s ← 0
for all x^μ ∈ A do
    s ← s + x^μ
end for
x̄ ← s/n
for all x^μ ∈ A do
    x̂^μ ← x^μ − x̄
    k ← 1
    while k ≤ C do
        if class(x̂^μ) = k then
            y_k^μ = 1
        else
            y_k^μ = 0
        end if
        k ← k + 1
    end while
end for
W ← 0
for all (x̂^μ, y^μ) do
    W ← W + (y^μ)(x̂^μ)^T
end for
```

### 2.2 Level 2: Feature selection

Let $E_{i,j}^\mu$ be the error of classifying a sample $\hat{\mathbf{x}}^\mu$ to class $i$ based on gene $j$ expressed as,

$$E_{i,j}^\mu = \begin{cases} 1 & \text{if } w_{i,j}\hat{x}_j^\mu y_i^\mu < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then the cumulative classification error for gene $j$ can be defined as,

$$E_j = \sum_{i=1}^{C}\sum_{\mu=1}^{n} E_{i,j}^\mu \tag{2}$$

and let $\Theta$ be a reference value calculated from the cumulative errors $E_j$ for $j = 1, 2, \ldots, G$,

$$\Theta = \frac{1}{G}\sum_{j=1}^{G}\left(1 - \frac{1}{G}E_j\right). \tag{3}$$

Now we can construct a $G$-dimensional vector $\mathbf{V}$ whose $j$-th element is obtained as follows:

$$v_j = \begin{cases} 1 & \text{if } \left(1 - \frac{1}{G}E_j\right) > \Theta \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that those components of vector $\mathbf{V}$ whose value is equal to 0 will correspond to genes that can be deemed as irrelevant.

### 2.3 Classification using the HAM model

Once the matrix $\mathbf{W}$ and the vector $\mathbf{V}$ have been constructed, the classification of a new input sample $\mathbf{x}$ comprises two steps: (i) to obtain $\hat{\mathbf{x}}$ as a result of translating $\mathbf{x}$ according to step 7 of Algorithm 1, and (ii) to apply the recall stage of lernmatrix so that $\mathbf{x}$ is assigned to some of the $C$ classes.

The recall stage of lernmatrix consists of determining the elements of the vector $\mathbf{y}^\mu$ associated with an input sample $\mathbf{x}^\mu$. The $i$-th element $y_i^\mu$ of the class vector $\mathbf{y}^\mu$ is computed by using the following bipolar output function:

$$y_i^\mu = \begin{cases} 1 & \text{if } \sum_{j=1}^{G} w_{i,j}(\hat{x}_j^\mu v_j) \\ & = \max_{h=1}^{C} \left[\sum_{j=1}^{G} w_{h,j}(\hat{x}_j^\mu v_j)\right] \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

If an input vector $\hat{\mathbf{x}}^\mu$ is assigned to class $k$, this function yields a $C$-dimensional output vector $\mathbf{y}^\mu$ whose $k$-th element $y_k^\mu$ is set to 1 and the rest of elements $y_j^\mu$ ($j = 1, 2, \ldots, k-1, k+1, \ldots, C$) are set to 0.

## 3 Experiments

Applicability and efficiency of the new model have been analyzed by conducting a pool of experiments on four gene expression databases, which have been gathered from a public repository available at http://datam.i2r.a-star.edu.sg/datasets/krbd. Table 1 provides a quantitative comparison of the databases used in the experiments, reporting the number of genes (features), the cardinality of the data set, the number of samples in each class, and the imbalance ratio (the size of the majority class divided by the size of the minority class). It also includes the T2 measure (it describes the density of spatial distributions of samples by comparing the number of samples in the data set to the number of genes) [20].

As can be observed, the properties of the data sets chosen for the experiments reflect the challenging 'large $G$, small $n$' problem mentioned in Sect. 1, with small sample size ranging from 39 to 62 and high dimensionality ranging from 2000 to 7129 genes. This problem can be better seen by the values of T2, which are all very low (ranging from 0.0084 in CNS to 0.0310 in Colon). Conversely, differences in size between both classes are quite small (less than two samples from the majority class per each sample from the minority class) and therefore it appears that class imbalance should not represent a critical problem for these databases.

The CNS data correspond to the outcome of a treatment for central nervous system embryonal tumor, where the survivors refer to individuals who are alive after receiving the prescribed treatment and the failures are patients who passed away. The Colon database consists of 62 samples collected from individuals with colorectal cancer; the samples can be tumor or normal tissues (gathered from healthy parts of the colons of the same individuals). The DLBCL-Standford database collects samples from two groups of patients according to the gene expression profiling of diffuse large B cell lymphoma: one group has gene expression characteristics of germinal center B cells, and the other group corresponds to genes normally induced during in vitro activation of peripheral blood B cells. Finally, the Lung-Ontario database includes gene expression microarrays of individuals with non-small cell lung cancer, indicating whether they experienced relapse of their tumor or they are disease free.

It has to be noted that both DLBCL-Standford and Lung-Ontario databases contain several features with missing values. Therefore, to apply the classification models, these data sets have been preprocessed using the K-means clustering technique for missing data imputation [22,32].

### 3.1 Experimental design

We compared the two-level hetero-associative memory (HAM) against five standard prediction models: the nearest neighbor (NN) classifier with the Euclidean distance function, a support vector machine (SVM), a multi-layer perceptron network (MLP), a C4.5 decision tree and a random forest (RF). The architectures and parameter settings of each model were defined as follows. The MLP neural network was designed with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm, two neurons in the hidden layer, a sigmoid transfer function, the backpropagation learning algorithm, a learning rate $\alpha = 0.1$ and a maximum number of training epochs equal to 10,000; the SVM employed a linear kernel because this has been deemed as one of the best performing functions in numerous biomedical applications [8], using a soft-margin constant equal to 1.0, a tolerance of 0.001, a round-off error $\varepsilon = 1.0\text{E}{-12}$ and the

**Table 1** Some characteristics of the databases

| | # Genes | # Samples | Class1–Class2 | Imbalance | T2 |
|---|---|---|---|---|---|
| CNS | 7129 | 60 | Failure (21)–survivor (39) | 1.86 | 0.0084 |
| Colon | 2000 | 62 | Tumor (40)–normal (22) | 1.82 | 0.0310 |
| DLBCL-Standford | 4026 | 47 | Germinal (24)–activated (23) | 1.04 | 0.0117 |
| Lung-Ontario | 2880 | 39 | Relapse (24)–no relapse (15) | 1.60 | 0.0135 |

sequential minimal optimization algorithm; the C4.5 classifier was applied with a pruning confidence factor of 0.25; and each bag in the RF contained all training samples and the number of iterations was equal to 100. We also included the associative memory without feature selection (ASM) [12] to gain a better understanding of the behavior of HAM.

The fivefold cross-validation method was used in this study because it seems to be more appropriate and statistically safe than other well-known techniques, such as holdout with the need of large data sets to achieve good generalization or bootstrapping with critical assumptions to be taken (e.g., independence of samples and large data size) [10,24]. Each original set was randomly split into five stratified portions of equal size; for each fold, four parts were merged to build a training set, and the remaining block was put aside to have an independent set for testing purposes. Then the final scores presented in next section will correspond to the average of the five trials.

### 3.2 Performance assessment criteria

In general, most biological and biomedical applications need to evaluate not only the overall accuracy, but also the true-positive and true-negative hits because of the asymmetric costs of false-positives and false-negatives [29,34]. Hence, the performance of the algorithms has been here evaluated using three measures calculated from a $2 \times 2$ confusion matrix, in which each element indicates the number of correct/incorrect classifications:

– Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \times 100$$

– Sensitivity: percentage of positive samples that are classified correctly

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

– Specificity: percentage of negative samples that are classified correctly

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100,$$

where TP and TN refer to the total amount of positive and negative samples correctly predicted, respectively, while FP and FN indicate the total number of mispredictions on negative and positive samples, respectively.

Note that in the present study, the samples that belong to Class1 have been considered to mold the positive class and the samples from Class2 have comprised the negative class.

**Table 2** Classification accuracy

| | CNS | Colon | DLBCL | Lung |
|---|---|---|---|---|
| HAM | **68.33** | 79.03 | **96.00** | 75.64 |
| NN | 56.67 | 75.81 | 74.47 | 58.97 |
| MLP | 65.00 | 69.35 | 93.62 | 58.97 |
| SVM | **68.33** | 80.65 | 95.74 | **79.49** |
| C4.5 | 55.00 | **83.87** | 70.21 | 74.36 |
| RF | 61.67 | 79.03 | 91.49 | 69.23 |
| ASM | 67.69 | 77.69 | 90.88 | 61.43 |

The best performing model for each database is in bold

## 4 Results

Table 2 reports the classification accuracy for each database using the six prediction algorithms. The best performing model for each database is shown in boldface. The results indicate that no method performed the best for all databases, which reveals the complexity of tumor classification because of the heterogeneity of gene expression microarray data.

In general, the SVM was superior to the rest of algorithms (in fact, this was already known in microarray literature), but the HAM approach performed equally well as SVM on the CNS, and even better on DLBCL-Standford. In the case of CNS, NN and C4.5 appear not to be appropriate choices because their accuracy was very close to the performance of the random-guess classifier (i.e., accuracy $\approx 50.00\%$). Similar comments can be drawn for the NN and MLP models on the Lung-Ontario database where their accuracy was 58.97%, and also for ASM with an accuracy of 61.43%. Regarding the Colon database, the C4.5 decision tree and MLP were the best (83.87%) and the worst (69.35%) algorithms, respectively, while there were no remarkable differences between HAM, NN, SVM, RF and ASM.

Table 3 summarizes some descriptive statistics of the accuracy results for a complete understanding of the differences between the six prediction models. It can be seen that the classifier with the highest average accuracy corresponds to SVM, closely followed by the HAM method. The ranges between the maximum and minimum values of these two algorithms were similar, revealing that differences in accuracy were not critical.

Since the above observations result subjective and qualitative in nature, statistical tests can provide more objective insights into whether there exist some statistically significant differences. Our first step was to determine whether the data are approximately normally distributed. To this end, we run Shapiro–Wilk, Lilliefors and Jarque–Bera tests because graphical methods such as the frequency distribution and Q–Q plots are not very useful when the sample size is small. Thus, the null and alternative hypotheses for the normality tests were:

**Table 3** Descriptive statistics of accuracy results

| | Range | Q1 | Q3 | Mean | Std. Dev. |
|---|---|---|---|---|---|
| HAM | 27.67 | 73.81 | 83.27 | 79.75 | 11.72 |
| NN | 19.14 | 58.40 | 74.81 | 66.48 | 10.06 |
| MLP | 37.35 | 63.49 | 76.09 | 72.41 | 16.50 |
| SVM | 27.41 | 76.70 | 84.42 | 81.05 | 11.26 |
| C4.5 | 28.87 | 66.41 | 76.74 | 70.86 | 12.02 |
| RF | 29.82 | 67.34 | 82.15 | 75.36 | 12.89 |
| ASM | 29.45 | 66.13 | 80.99 | 74.42 | 12.85 |

**Table 4** Normality tests

| | Shapiro–Wilk | Lilliefors | Jarque–Bera |
|---|---|---|---|
| HAM | 0.599 | 0.356 | 0.805 |
| NN | 0.133 | 0.293 | 0.727 |
| MLP | 0.236 | 0.147 | 0.710 |
| SVM | 0.732 | 0.416 | 0.895 |
| C4.5 | 0.857 | 0.651 | 0.869 |
| RF | 0.912 | 0.909 | 0.845 |
| ASM | 0.896 | 0.901 | 0.835 |

- **H0**: The data follow a normal distribution.
- **H1**: The data do not follow a normal distribution.

Results of Shapiro–Wilk, Lilliefors and Jarque–Bera tests are shown in Table 4. All normality tests failed to reject the null hypothesis ($p$ value $< 0.05$), which means that there is not enough evidence to conclude that the data do not follow a normal distribution.

As the normality assumption was not rejected, we performed paired sample $t$ tests between the set of results corresponding to each pair of methods. Thus, we checked whether the mean of accuracies of one model $M_i$ was more significant than the mean of accuracies of another model $M_j$ at the 95% confidence level. In this case, the null and alternative hypotheses for the $t$ tests were:

- **H0**: The difference between the means is equal to 0.
- **H1**: The difference between the means is different from 0.

We rejected the null hypothesis H0 and accepted the alternative hypothesis H1 if the $p$ value was less than 0.05.

The numbers marked as bold in Table 5 indicate that the paired sample $t$ test rejected the null hypothesis ($p$ value $< 0.05$). It can be observed that the only cases with statistically significant differences corresponded to the test between HAM and NN and the test between SVM and NN. Nevertheless, the risk to reject the null hypothesis H0 while it is true was low for the comparisons of HAM and SVM with the rest of models.

**Table 5** Paired sample $t$ tests

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| (1) HAM | | **0.04** | 0.14 | 0.26 | 0.28 | 0.07 | 0.19 |
| (2) NN | | | 0.40 | **0.03** | 0.41 | 0.06 | 0.11 |
| (3) MLP | | | | 0.19 | 0.89 | 0.52 | 0.53 |
| (4) SVM | | | | | 0.19 | 0.05 | 0.19 |
| (5) C4.5 | | | | | | 0.52 | 0.68 |
| (6) RF | | | | | | | 0.76 |
| (7) ASM | | | | | | | |

The test rejected the null hypothesis is in bold

**Table 6** Genes selected using the HAM model

| | CNS | Colon | DLBCL | Lung |
|---|---|---|---|---|
| # Genes | 3882 | 932 | 1727 | 1572 |
| % Reduction | 45.54 | 53.40 | 57.10 | 45.42 |

To analyze the differences between the two best performing algorithms (HAM and SVM), we computed four data complexity measures [20] that allow to quantify the overlap between classes (F1 and F3) and the class separability (L1 and L2). Thus, we found that Colon constitutes a more complex problem than DLBCL-Standford because the values of F1 and F3 were 1.083 and 0.000 for Colon, and 2.907 and 0.016 for DLBCL-Standford. Analogously, the values of L1 and L2 were 0.498 and 0.000 for Colon, and 0.141 and 0.032 for DLBCL-Standford. The reasons why we put our attention on the accuracies for Colon and DLBCL-Standford are twofold: (i) the highest differences between SVM and HAM were achieved for the Colon database and (ii) HAM outperformed SVM when applied to the DLBCL-Standford data set.

Table 6 gives the number of genes selected by the HAM method and the percentage of reduction over the total amount of genes for each database. It is worth remarking that the use of the hybrid hetero-associative memory proposed here may enable significant gains in computational speed and memory because the irrelevant and/or redundant genes are removed from the data sets. It was found that the percentage of reduction was about 45–55% in average, which represents a quite meaningful amount when the number of genes is very large (over thousands of genes) as is the case of the microarray data sets.

The use of accuracy to assess the prediction performance is not the most suitable metric in real-life problems with skewed class distributions and unequal misclassification costs [25], as is the case of tumor (cancer) classification/prediction. Besides, this is also related to the likelihood of false-positives

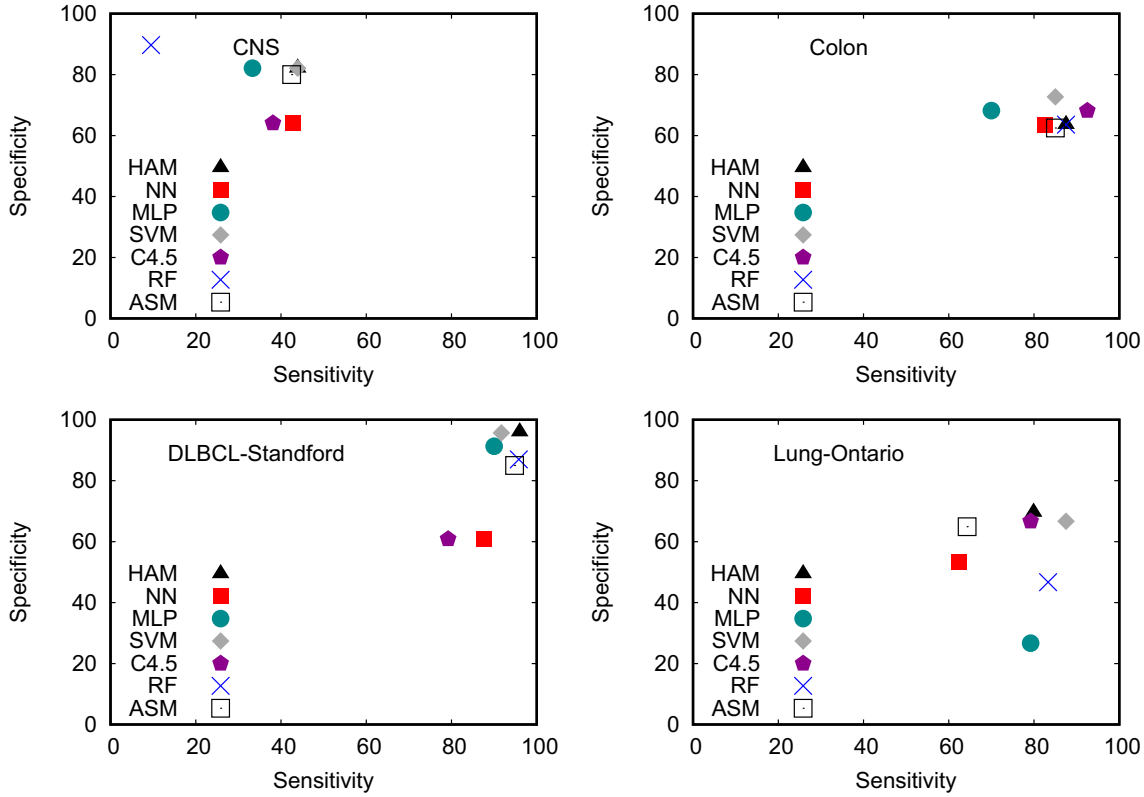


**Fig. 1** Plots of sensitivity versus specificity for each database

as a result of the 'large $G$, small $n$' problem in gene expression databases [15]. Hence, we have also included the values of sensitivity and specificity to avoid potentially misleading conclusions drawn from accuracy since this tends to be heavily biased toward the majority class.

Figure 1 displays the classification procedures in a space where the $x$-axis shows the sensitivity and the $y$-axis depicts the specificity. In such a space, a model with perfect prediction will be placed on the upper right corner (100% sensitivity, 100% specificity) of the plot. Therefore, the closer the method is to the upper right corner, the higher the classification performance on both classes. Nevertheless, for most biomedical applications it is better not to miss a diagnosis of disease rather than to err in the classification of a non-tumor sample, which suggests that maximizing the sensitivity (points close to the right side) is of far more important than the specificity (points close to the upper side).

One can observe that both HAM and SVM were the algorithms with the most balanced trade-off between sensitivity and specificity, whereas the behavior of the remaining models depended on each particular database, especially in the case of MLP and RF. For instance, MLP applied to Lung-Ontario data achieved about 80% of sensitivity and 25% of specificity, which reveals that it misclassified many negative samples. On the other hand, RF is located close to the upper left corner of the plot (very low sensitivity and high specificity), indicating that this model failed on the prediction of most positive samples (the most important cases), and therefore, this classifier is of no value at all for this specific problem. Finally, note that the associative memory without feature selection also showed a right balance between sensitivity and specificity.

## 5 Conclusions

This paper has introduced a two-level algorithm for tumor classification and characterization from gene expression microarray data. The proposed technique comprises two stages: the first one aims to construct a particular type of hetero-associative memory and the second level allows for the selection of the most differentially expressed genes. The neural network here designed corresponds to a combination of the linear associator and the Steinbuch's lernmatrix, and it also includes an initial step in which the coordinate axes are firstly translated to a new origin.

The HAM prediction model has been evaluated on four gene expression microarray databases and empirically compared to five well-established classifiers (SVM, MLP, NN, C4.5 and RF) and also to the associative memory without

feature selection (ASM) by measuring three scores: overall accuracy, sensitivity and specificity. The results have shown that the two-level hetero-associative memory has performed similar to the best prediction model in most cases, but it has also been observed that these comparisons strongly depend on the particular characteristics of each database. However, importantly for practical applications, an attractive advantage of the HAM approach here introduced is the significant reduction in the number of genes used for classification, which may lead to a considerable decrease in computational requirements and help to increase its biological interpretability.

The method proposed in this paper represents a meaningful contribution to the collection of strategies for the classification, characterization and analysis of gene expression microarrays in cancer research. However, it still constitutes a first step toward exploring more complex hybridization techniques that merge classification and feature selection through associative neural networks or other emerging connectionist models (e.g., deep neural networks), thus providing a better understanding of medical decisions because a lower number of genes should be analyzed. Another direction for further research is to consider the HAM algorithm as a feature selection method for other classifiers.

**Acknowledgements** This study was partially supported by the Valencian Council of Education, Research, Culture and Sport [PROMETEOII/2014/062], the Mexican PRODEP [DSA/103.5/15/7004], and the Spanish Ministry of Economy, Industry and Competitiveness under Grant [TIN2013-46522-P].

## References

1. Aghajari, Z.H., Teshnehlab, M., Jahed Motlagh, M.R.: A novel chaotic hetero-associative memory. Neurocomputing **167**, 352–358 (2015)
2. Aihara, K., Takabe, T., Toyoda, M.: Chaotic neural networks. Phys. Lett. A **144**(6), 333–340 (1990)
3. Aldape-Pérez, M., Yáñez-Márquez, C., Camacho-Nieto, O., Argüelles-Cruz, A.J.: An associative memory approach to medical decision support systems. Comput. Methods Prog. Biomed. **106**(3), 287–307 (2012)
4. Anderson, J.A.: A simple neural network generating an interactive memory. Math. Biosci. **14**, 197–220 (1972)
5. Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A.: Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE ACM Trans Comput. Biol. Bioinform. **13**(5), 971–989 (2016)
6. Arya, K.V., Singh, V., Mitra, P., Gupta, P.: Face recognition using parallel associative memory. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Singapore, pp. 1332–1336 (2008)
7. Babu, M., Sarkar, K.: A comparative study of gene selection methods for cancer classification using microarray data. In: Proceedings of the 2nd International Conference on Research in Computational Intelligence and Communication Networks, Kolkata, India, pp. 204–211 (2016)
8. Ben-Hur, A., Weston, J.: A user's guide to support vector machines. In: Carugo, O., Eisenhaber, F. (eds.) Data Mining Techniques for the Life Sciences, Methods in Molecular Biology, vol. 609, pp. 223–239. Humana Press, New York (2010)
9. Berns, A.: Cancer: gene expression in diagnosis. Nature **403**, 491–492 (2000)
10. Braga-Neto, U.M., Dougherty, E.R.: Is cross-validation valid for small-sample microarray classification? Bioinformatics **20**(3), 374–380 (2004)
11. Chartier, S., Lepage, R.: Learning and extracting edges from images by a modified hopfield neural network. In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec City, Canada, vol. 3, pp. 431–434 (2002)
12. Cleofas-Sánchez, L., García, V., Marqués, A., Sánchez, J.: Financial distress prediction using the hybrid associative memory with translation. Appl. Soft Comput. **44**, 144–152 (2016)
13. Dougherty, E.R.: Small sample issues for microarray-based classification. Comp. Funct. Genom. **2**(1), 28–34 (2001)
14. Dudoit, S., Fridlyand, J.: Classification in microarray experiments. In: Speed, T.P. (ed.) Statistical Analysis of Gene Expression Microarray Data, pp. 93–158. Chapman & Hall/CRC Press, London (2003)
15. Ein-Dor, L., Zuk, O., Domany, E.: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. **103**(15), 5923–5928 (2006)
16. García, V., Sánchez, J.S.: Mapping microarray gene expression data into dissimilarity spaces for tumor classification. Inform. Sci. **294**, 362–375 (2015)
17. García, V., Sánchez, J.S., Cleofas-Sánchez, L., Ochoa-Domínguez, H.J., López-Orozco, F.: An insight on the 'large G, small n' problem in gene-expression microarray classification. In: Proceedings of the 8th Iberian Conference on Pattern Recognition and Image Analysis, Faro, Portugal, pp. 483–490 (2017)
18. Hassanien, A.E., Al-Shammari, E.T., Ghali, N.I.: Computational intelligence techniques in bioinformatics. Comput. Biol. Chem. **47**, 37–47 (2013)
19. Hira, Z.M., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinform. **2015**(ID 198363), 1–13 (2015)
20. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 289–300 (2002)
21. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. In: Anderson, J.A., Rosenfeld, E. (eds.) Neurocomputing: Foundations of Research, pp. 457–464. Proceedings of the National Academy of Sciences USA, Cambridge (1988)
22. Hruschka, E.R., Hruschka, E.R., Ebecken, N.F.F.: Towards efficient imputation by nearest-neighbors: a clustering-based approach. In: Proceedings of the 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, pp. 513–525 (2004)
23. Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R.: Optimal number of features as a function of sample size for various classification rules. Bioinformatics **21**(8), 1509–1515 (2005)
24. Irsoy, O., Yildiz, O.T., Alpaydin, E.: Design and analysis of classifier learning experiments in bioinformatics: survey and case studies. IEEE ACM Trans. Comput. Biol. **9**(6), 1663–1675 (2012)
25. Japkowicz, N.: Assessment metrics for imbalanced learning. In: He, H., Ma, Y. (eds.) Imbalanced Learning: Foundations, Algorithms, and Applications, pp. 187–210. Wiley IEEE Press, New York (2013)
26. Kohonen, T.: Correlation matrix memories. IEEE Trans. Comput. **C–21**(4), 353–359 (1972)
27. Kohonen, T.: Associative Memory. A System—Theoretical Approach. Springer, Berlin (1977)

28. Kosko, B.: Bidirectional associative memories. IEEE Trans. Syst. Man Cybern. **18**(1), 49–60 (1988)
29. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A., Robles, V.: Machine learning in bioinformatics. Brief. Bioinform. **7**(1), 86–112 (2011)
30. 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)
31. Lee, J.W., Lee, J.B., Park, M., Song, S.H.: An extensive evaluation of recent classification tools applied to microarray data. Comput. Stat. Data Anal. **48**, 869–885 (2005)
32. Li, D., Deogun, J., Spaulding, W., Shuart, B.: Towards missing data imputation: a study of fuzzy K-means clustering method. In: Proceedings of the 4th International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden, pp. 573–579 (2004)
33. Lu, Y., Han, J.: Cancer classification using gene expression data. Inform. Syst. **28**(4), 243–268 (2003)
34. Ma, S., Huang, J.: Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics **21**(2), 4356–4362 (2005)
35. Mahata, P., Mahata, K.: Selecting differentially expressed genes using minimum probability of classification error. J. Biomed. Inform. **40**(6), 775–786 (2007)
36. Nakano, K.: Associatron—a model on associative memory. IEEE Trans. Syst. Man Cybern. **2**(3), 380–388 (1972)
37. 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)
38. Raudys, S.J., Jain, A.K.: Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. **13**(3), 252–264 (1991)
39. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
40. Sharma, N., Ray, A., Sharma, S., Shukla, K., Pradhan, S., Aggarwal, L.: Segmentation and classification of medical images using texture-primitive features: application of BAM-type artificial neural network. J. Med. Phys. **33**(3), 119–126 (2008)
41. Steinbuch, K.: Die lernmatrix. Kybernetik **1**(1), 36–45 (1961). In German
42. Sudo, A., Sato, A., Hasegawa, O.: Associative memory for online learning in noisy environments using self-organizing incremental neural network. IEEE Trans. Neural Netw. **20**(6), 964–972 (2009)
43. Sun, X., Liu, Y., Wei, D., Xu, M., Chen, H., Han, J.: Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis. J. Biomed. Inform. **46**(2), 252–258 (2013)
44. Vaishnavi, Y., Shreyas, R., Suhas, S., Surya, U.N., Ladwani, V.M., Ramasubramanian, V.: Associative memory framework for speech recognition: adaptation of hopfield network. In: 2016 IEEE Annual India Conference, Bangalore, India, pp. 1–6 (2016)
45. Villuendas-Rey, Y., Rey-Benguría, C.F., Ferreira-Santiago, A., Camacho-Nieto, O., Yáñez-Márquez, C.: The naïve associative classifier (NAC): a novel, simple, transparent, and accurate classification model evaluated on financial data. Neurocomputing **265**, 105–115 (2017)
46. Weigelt, B., Baehner, F.L., Reis-Filho, J.S.: The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J. Pathol. **220**(2), 263–280 (2010)
47. Xing, E.P., Jordan, M.I., Karp, R.M.: Feature selection for high-dimensional genomic microarray data. In: Proceedings of the 8th International Conference on Machine Learning, Williamstown, MA, pp. 601–608 (2001)
48. Yáñez-Márquez, C.: Associative memories based on order relations and binary operators. Ph.D. thesis, Centro de Investigación en Computación - Instituto Politécnico Nacional, Mexico, **(In Spanish)** (2002)
49. Yoon, Y., Lee, J., Park, S., Bien, S., Chung, H.C., Rha, S.Y.: Direct integration of microarrays for selecting informative genes and phenotype classification. Inf. Sci. **178**(1), 88–105 (2008)
50. Zhang, Z., Zhuo, H., Liu, S., de B Harrington, P.: Classification of cancer patients based on elemental contents of serums using bidirectional associative memory networks. Anal. Chim. Acta **436**(2), 281–291 (2001)