Feature Selection for Microarray Gene Expression Data Using Simulated Annealing Guided by the Multivariate Joint Entropy

Félix Fernando González-Navarro[1] and Lluís A. Belanche-Muñoz[2]

[1] Instituto de Ingeniería,
Universidad Autónoma de Baja California, Mexicali, Mexico

[2] Dept. de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya, Barcelona, Spain

fernando.gonzalez@uabc.edu.mx, belanche@lsi.upc.edu

Abstract. Microarray classification poses many challenges for data analysis, given that a gene expression data set may consist of dozens of observations with thousands or even tens of thousands of genes. In this context, *feature subset selection* techniques can be very useful to reduce the representation space to one that is manageable by classification techniques. In this work we use the discretized multivariate *joint entropy* as the basis for a fast evaluation of gene relevance in a Microarray Gene Expression context. The proposed algorithm combines a *simulated annealing* schedule specially designed for feature subset selection with the incrementally computed joint entropy, reusing previous values to compute current feature subset relevance. This combination turns out to be a powerful tool when applied to the maximization of gene subset relevance. Our method delivers highly interpretable solutions that are more accurate than competing methods. The algorithm is fast, effective and has no critical parameters. The experimental results in several public-domain microarray data sets show a notoriously high classification performance and low size subsets, formed mostly by biologically meaningful genes. The technique is general and could be used in other similar scenarios.

Keywords. Feature selection, microarray gene expression data, multivariate joint entropy, simulated annealing.

Selección de características para datos de expresión de los genes mediante microarreglos usando recocido simulado guiado por la entropía conjunta multivariada

Resumen. La clasificación de microarreglos plantea muchos desafíos para el análisis de datos, dado que un conjunto de datos de expresión de genes puede contener docenas de observaciones con miles o incluso decenas de miles de genes. En este contexto, las técnicas de selección de subconjuntos de características pueden ser muy útiles para reducir el espacio de representación a uno manejable mediante técnicas de clasificación. En este trabajo se utiliza la entropía conjunta discretizada multivariada como base para la evaluación rápida de la relevancia de genes en el contexto de expresión génica mediante microarreglos. El algoritmo propuesto desarrolla una técnica de recocido simulado diseñada especialmente para la selección de subconjuntos de características, a través de la entropía conjunta. Ésta es calculada incrementalmente, reutilizando los valores anteriores para calcular la relevancia de los subconjuntos de características. Esta combinación resulta ser una herramienta poderosa cuando se aplica a la maximización de la relevancia de un subconjunto de genes. Nuestro método ofrece soluciones altamente interpretables y más precisas que las propuestas por métodos competidores. El algoritmo propuesto es rápido, eficaz y no presenta parámetros críticos. Los resultados de los experimentos con varios conjuntos de datos de microarreglos de dominio público revelan alto rendimiento de clasificacíń y subconjuntos de pequeño tamaño, formados en su mayoría por genes biológicamente significativos. La técnica es general y podría ser utilizada en otros escenarios similares.

Palabras clave. Selección de características, datos de expresiones de los genes mediante microarreglos, entropía conjunta multivariada, recocido simulado.

Computación y Sistemas Vol. 18 No. 2, 2014 pp. 275-293
ISSN 1405-5546
http://dx.doi.org/10.13053/CyS-18-2-2014-032

1 Introduction

In cancer diagnosis, accurate classification of the different tumor types is of paramount importance. An accurate prediction of different tumor types provides better treatment and toxicity minimization on patients. Traditional methods of tackling this situation are based primarily on morphological characteristics of tumorous tissue [13]. These conventional methods are reported to have several diagnosis limitations. In order to analyze the problem of cancer classification using gene expression data, more systematic approaches were developed [34].

Pioneering work in cancer classification by gene expression using DNA microarray showed the possibility to help the diagnosis by means of Machine Learning or more general Data Mining methods [22], which are now extensively used for this task [16]. However, in this setting gene expression data analysis entails a heavy computational consumption of resources, due to the extreme sparseness compared to standard data sets in classification tasks [51]. Classifying cancer types using such a very high ratio of the number of variables to the number of observations is a delicate process. As a result, dimensionality reduction and in particular *feature subset selection* (FSS) techniques may be very useful. The finding of small subsets of very relevant genes among a huge quantity of genes could result in much specific and thus efficient treatments.

This work addresses the problem of selecting a subset of features by using the TAFS (Thermodynamic Algorithms for Feature Selection) family of methods for the FSS problem. Given a suitable objective function, these algorithms make use of a special-purpose *simulated annealing* (SA) technique to find a good subset of features that maximizes the objective function. A distinctive characteristic over other search algorithms for FSS is the probabilistic capability to momentarily accept worse solutions, which in the end may result in better hypotheses.

Despite their powerful optimization capability, SA-based search algorithms usually lack execution speed, involving long convergence times. In consequence, they have been generally excluded as an option in FSS problems, let alone in highly complex domains such as microarray gene expression data. Nonetheless, a few contributions using the classical SA algorithm for FSS are found in prostate protein mass spectrometry data [32], marketing applications [35], or parameter optimization in clustering gene expression analysis [18].

Our answer to these computational problems is twofold. First, we use a *filter* objective function for FSS (thus avoiding the development of a predictive model for every subset evaluation). Second, the objective function itself is evaluated very efficiently based in the reutilization of previous computations. Specifically, a way to calculate the multivariate joint entropy for categorical variables is presented that is both exact and very efficient. This measure is then used by a SA-based TAFS algorithm to search for small subsets of highly relevant genes. Classification experiments in five public domain microarray datasets yield some of the best prediction results reported so far for these problems while offering a drastic reduction in subset sizes.

The paper is organized as follows. Section 2 briefly reviews the necessary background. Section 3 develops the proposed method, the new information-theoretic measure for feature relevance, its efficient implementation and its embedding into a TAFS-like algorithm, which we name μ-TAFS. Section 4 describes the data sets and the experimental settings, Section 5 presents the results and their interpretation. The paper ends with the conclusions and directions for future work.

2 Preliminaries

In this section we briefly review the necessary background: the Simulated Annealing technique, basic Information Theory concepts and the TAFS family of thermodynamic algorithms for feature subset selection.

2.1 Simulated Annealing

Simulated Annealing (SA) is a stochastic technique inspired on statistical mechanics for finding (near) globally optimal solutions to large optimization problems. SA is a weak method in that it needs almost no information about the structure of the search space. The algorithm works by assuming that some parts of the current solution belong to a potentially better one, and thus these parts should be retained by exploring neighbors of the current solution. Assuming the objective function is to be minimized, then SA would jump from hill to hill and hence escape or simply avoid sub-optimal solutions.

When a system S (considered as a set of possible states) is in thermal equilibrium (at a given temperature T), the probability that it is in a certain state s, called $P_T(s)$, depends on T and on the energy $E(s)$ of the state s. This probability follows a Boltzmann distribution:

$$P_T(s) = \frac{\exp\left(-\frac{E(s)}{kT}\right)}{Z},\ with\ \ Z = \sum_{s\in S} \exp\left(-\frac{E(s)}{kT}\right)$$

where k is the Boltzmann constant and Z acts as a normalization factor. Metropolis and his coworkers developed a stochastic relaxation method that works by simulating the behavior of a system at a given temperature T [36]. Being s the current state and s' a neighboring state, the probability of making a transition from s to s' is the ratio $P_T(s \to s')$ of the probability of being in s to the probability of being in s':

$$P_T(s \to s') = \frac{P_T(s')}{P_T(s)} = \exp\left(-\frac{\Delta E}{kT}\right) \qquad (1)$$

where we have defined $\Delta E = E(s') - E(s)$. Therefore, the acceptance or rejection of s' as the new state depends on the difference of the energies of both states at temperature T. If $P_T(s') \geq P_T(s)$ then the “move” is always accepted. It $P_T(s') < P_T(s)$ then it is accepted with probability $P_T(s, s') < 1$ (this situation corresponds to a transition to a higher-energy state).

Note that this probability depends upon the current temperature T and decreases as T does. In the end, there will be a value of T low enough (the *freezing point*), wherein these transitions will be very unlikely and the system will be considered frozen. In order to maximize the probability of finding states of minimal energy at every value of T, *thermal equilibrium* must be reached. To do this, according to Metropolis, an annealing schedule is designed to prevent the process from getting stuck at a local minimum. The SA algorithm introduced in [30] consists in using the Metropolis idea at each temperature T for a finite amount of time. In this algorithm T is first set at a initially high value, spending enough time at it in order to approximate thermal equilibrium. Then a small decrement of T is performed and the process is iterated until the system is considered frozen.

If the cooling schedule is well designed, the final reached state may be considered a near-optimal solution. However, the whole process is inherently slow, mainly because of the thermal equilibrium requirement at every temperature T.

2.2 Information Theory

Entropy, a main concept in Information Theory [47], can be seen as an average of the uncertainty in a random variable. If X is a discrete random variable with probability mass function (PMF) p, its entropy is defined by[1]

$$H(X) = -\sum_{x} p(x)\, \log p(x) = -E_X[\log\, p(X)] \quad (2)$$

being $E[\,]$ the expectation operator over the PMF of X. If a variable (X) is known and another one (Y) is not, the *conditional entropy* of Y with respect to X is the mutual entropy with respect to the corresponding conditional distribution:

$$H(Y|X) = -\sum_{x}\sum_{y} p(x,y) \log\, p(y|x). \qquad (3)$$

The *mutual information* (MI) can be interpreted as a measure of the information that a random variable has or explains about another one:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= E_{X,Y}[log\, \tfrac{p(x,y)}{p(x)p(y)}] \end{aligned} \qquad (4)$$

where H denotes the *entropy*. Note that $I(X;X) = H(X)$, since $H(X|X) = 0$ and $I(X;Y) = I(Y;X)$.

[1] All logarithms are taken in base 2.

The *conditional* MI is expressed in the natural way, by conditioning in (4):

$$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z) \tag{5}$$

MI has been successfully used in feature selection, as a way to measure the influence that a feature has over the class or target variable. Sometimes a normalized variant is used, given by $C_{XY} = \frac{I(X;Y)}{H(Y)}$, where Y is the class or target variable, commonly known as *coefficient of constraint* or *uncertainty coefficient* –note that the maximum value that $I(X;Y)$ can take is $H(Y)$. This coefficient can be understood by analyzing Fig. 1.

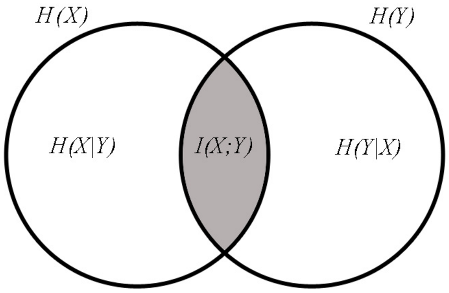


Fig. 1. Mutual information (eq. 4) between X and Y

Increasing the value $I(X;Y)$, $H(Y|X)$ is decreased which means that the uncertainty about a variable Y *is reduced* by the knowledge of X. There is an index of relevance that exploits this property to measure the relevance of a feature X (with respect to a class or target value Y) by conditioning on a third variable Z [4]:

$$\begin{aligned} R(X;Y|Z) &= \frac{I(X;Y|Z)}{H(Y|Z)} \\ &= \frac{H(Y|Z) - H(Y|X,Z)}{H(Y|Z)} \end{aligned} \tag{6}$$

where $R(X;Y|Z) = 0$ if $H(Y|Z) = 0$. Using a forward selection strategy to maximize it, Eq. (6) has been applied with some success to low-dimensional data sets [4].

2.3 Thermodynamic Algorithms for Feature Selection

In this section we review TAFS (Thermodynamic Algorithm for Feature Selection), an algorithm for FSS that was originally designed for problems of moderate feature size (up to one hundred) [23]. If we consider FSS as a search of possible feature subsets of the full feature set $\mathcal{X}$, then SA acts as a combinatorial optimization process [41]. In this sense, TAFS finds a subset of features that optimize the value of a given objective function $J : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, which we assume as non-negative and to be maximized[2].

To this end, a special-purpose forward/backward mechanism is embedded into an SA algorithm, taking advantage of its most distinctive characteristic: the probabilistic acceptance of worse scenarios over a finite time. This characteristic is enhanced by the notion of an ϵ-improvement: a feature ϵ-improves a current solution if it has a higher value of the objective function or a value not worse than ϵ%. This mechanism is intended to account for noise in the evaluation of the objective function (caused either by the finiteness of the data set or introduced by the chosen resampling method).

The pseudo-code of TAFS is presented in **Algorithm 1**. The algorithm consists of two major loops. The outer loop waits for the inner loop to finish and then updates T according to the chosen cooling schedule. When this loop reaches T_{min}, the algorithm halts. It keeps track of the best solution found (which is not necessarily the current one).

The inner loop is the core of the algorithm and is composed of two interleaved procedures: *Forward* and *Backward*, that iterate until an equilibrium point is found. These procedures work independently of each other, but share information about the *results* of their respective searches (the current solution). Within each of them, FSS takes place and the mechanism to escape from local minima starts working. The pseudo-code for Forward and Backward procedures, and ϵ-improvement is outlined in **Algorithms 2**, **3** and **4**. These procedures iteratively add or remove features one at a time in

[2] This is the case of accuracy, mutual information, inter-class distances and many other useful measures.

Algorithm 1: TAFS algorithm for feature selection

```
input : $\mathcal{X}$ : Full Feature set $\{X_1 \dots X_n\}$
        $J()$ : Objective Function
        $\alpha()$ : Cooling Schedule
        $\epsilon$ : Epsilon
        $T_0$ : Initial Temperature
        $T_{min}$ : Final Temperature
$X_{cur} \leftarrow \emptyset$ Initial Current Subset
$J_{cur} \leftarrow 0$ Initial Objective Function Value
$T \leftarrow T_0$ Initial Temperature
while $T > T_{min}$ do
    repeat
        $Y \leftarrow X_{cur}$
        $Forward\ (X_{cur}, J_{cur})$
        $Backward\ (X_{cur}, J_{cur})$
    until $Y = X_{cur}$
    $T \leftarrow \alpha(T)$
```

Algorithm 2: Procedure Forward (Z, J_Z are modified)

```
input : $Z, J_Z$
repeat
    $x \leftarrow \operatorname{argmax}_{X_i \in \mathcal{X} \setminus Z} J(Z \cup \{X_i\})$
    if $>_\epsilon (Z, x, true)$ then
        $accept \leftarrow true$
    else
        $\Delta J \leftarrow J(Z \cup \{x\}) - J(Z)$
        $accept \leftarrow rand(0,1) < e^{\frac{\Delta J}{T}}$
    if $accept$ then
        $Z \leftarrow Z \cup \{x\}$
    if $J(Z) > J_{cur}$ then
        $J_Z \leftarrow J(Z)$
until $not\ accept$
```

such a way that an ϵ-improvement is accepted unconditionally, whereas a non ϵ-improvement is accepted probabilistically. When *Forward* and *Backward* finish their respective tasks, TAFS checks if the current solution is the same as it was prior to their execution. If this is the case, then we consider that thermal equilibrium has been reached and T is adjusted, according to the cooling schedule. If it is not, another loop of *Forward* and *Backward* is carried out.

Algorithm 3: Procedure Backward (Z, J_Z are modified)

```
input : $Z, J_Z$
repeat
    $x \leftarrow \operatorname{argmax}_{X_i \in Z} J(Z \setminus \{X_i\})$
    if $>_\epsilon (Z, x, false)$ then
        $accept \leftarrow true$
    else
        $\Delta J \leftarrow J(Z \setminus \{x\}) - J(Z)$
        $accept \leftarrow rand(0,1) < e^{\frac{\Delta J}{T}}$
    if $accept$ then
        $Z \leftarrow Z \setminus \{x\}$
    if $J(Z) > J_Z$ then
        $J_Z \leftarrow J(Z)$
until $not\ accept$
```

Algorithm 4: Function $>_\epsilon$

```
input : $Z, x, d$
output: boolean
if $d$ then
    $Z' \leftarrow Z \cup \{x\}$
else
    $Z' \leftarrow Z \setminus \{x\}$
$\Delta x \leftarrow J(Z') - J(Z)$
if $\Delta x > 0$ then
    $return\ true$
else
    $return\ \frac{-\Delta x}{J(Z)} < \epsilon$
```

3 Proposed Method

The proposed method represents a development of the TAFS algorithm in three ways: first, we enhance the algorithm to make it much faster and effective; second, we derive a new information-theoretic measure for feature relevance; and third, we present an efficient incremental implementation of this measure. We name this new algorithm μ-TAFS.

3.1 eTAFS: an Enhanced TAFS Algorithm

A modification to Algorithm 1 aimed at speeding up relaxation time is presented in this section. The algorithm—named *e*TAFS, see **Algorithms 5** and **6**—is endowed with a *feature search window* (of size l) in the backward step as follows. In *forward* steps always the *best* feature is added (by looking at all possible additions). In *backward* steps this search is limited to l tries at random (without replacement). The value of l is incremented by one at every thermal-equilibrium point. This mechanism is an additional source of non-determinism and a bias towards adding a feature only when it is the best option available. On the contrary, to remove one, it suffices that its removal ϵ-improves the current solution. Another direct consequence is of course a considerable speed-up of the algorithm. Note that the design of *e*TAFS is such that it grows more and more deterministic, informed and costly as it converges toward the final configuration.

Algorithm 5: *e*TAFS algorithm for feature selection

```
input : $\mathcal{X}$ : Full Feature Set $\{X_1 \ldots X_n\}$
        $J()$ : Objective Function
        $\alpha()$ : Cooling Schedule
        $\epsilon$ : Epsilon
        $T_0$ Initial Temperature
        $T_{min}$ Final Temperature
$X_{cur} \leftarrow \emptyset$ Initial Current Subset
$J_{cur} \leftarrow 0$ Initial Objective Function Value
$T \leftarrow T_0$ Initial Temperature
$l \leftarrow 2$ Window Size (for backward steps)
while $T > T_{min}$ do
    repeat
        $Y \leftarrow X_{cur}$
        $Forward\ (X_{cur}, J_{cur}, l)$
        $Backward\ (X_{cur}, J_{cur}, l)$
    until $Y = X_{cur}$
    $T \leftarrow \alpha(T)$
    $l \leftarrow l + 1$
```

3.2 Information-Theoretic Feature Relevance

The definition in Eq. (6) lacks some components, Fig. 2 represents a three-variable interaction diagram. The stronger shaded area represents Eq. (6), where two components are missing, $I(X;Y;Z)$ and $I(Y;Z|X)$, and therefore normalization is done in a bigger area than it should, namely, in $H(Y|Z)$ rather than in $H(Y|X,Z)$. As a consequence, establishing a measure with respect to the reference entropy $H(Y)$ is incomplete.

In this work we calculate the conditional MI between a class variable Y and two variables X and Z as the joint information that X, Z explain about Y.

Algorithm 6: *e*TAFS Backward procedure (Z, J_Z are modified). Note that X_0 can be efficiently computed while in the **for** loop)

```
input : $Z, J_Z, l$
$A \leftarrow \emptyset; A_B \leftarrow \emptyset$
repeat
    for $i := 1$ to $min(l, |Z|)$ do
        Select $x \in Z \setminus A_B$ randomly
        if $>_\epsilon (Z, x, false)$ then
            $A \leftarrow A \cup \{x\}$
        $A_B \leftarrow A_B \cup \{x\}$
    $X_0 \leftarrow \underset{X \in A_B}{\mathrm{argmax}} \{J(Z \setminus \{X\})\}$
    if $X_0 \in A$ then
        $accept \leftarrow true$
    else
        $\Delta J \leftarrow J(Z \setminus \{X_0\}) - J(Z)$
        $accept \leftarrow rand(0,1) < e^{\frac{\Delta J}{t}}$
    if $accept$ then
        $Z \leftarrow Z \setminus \{X_0\}$
    if $J(Z) > J_Z$ then
        $J_Z \leftarrow J(Z)$
until $not\ accept$
```

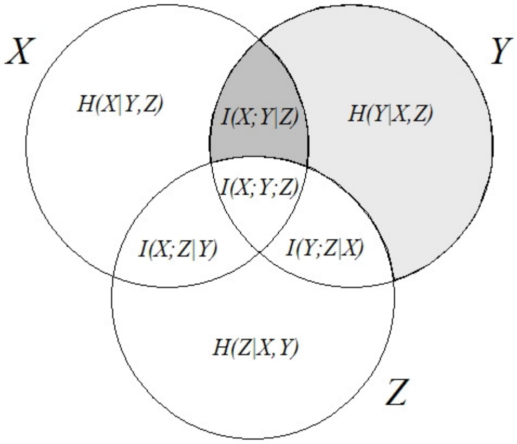


Fig. 2. Conditional mutual information (eq. 5) between X, Y and Z

The shaded area in Fig. 3 represents $I(Y;X,Z) = H(Y) - H(Y|X,Z)$, the information that two variables explain about a third one. Remarkably, this quantity can be calculated without explicit conditioning as follows:

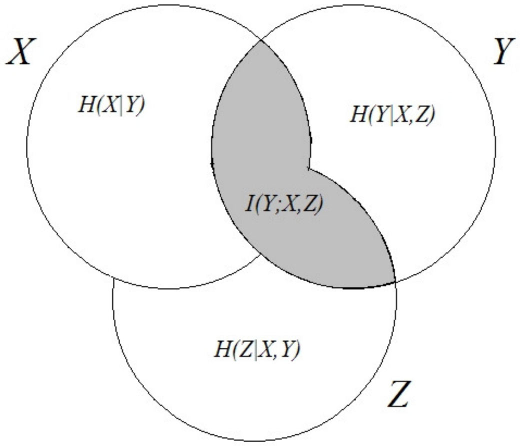


Fig. 3. Mutual information between two variables X and Z and a class variable Y, as in Eq. (7)

$$
\begin{aligned}
I(Y;X,Z) = {} & H(Y) - H(Y|X,Z) \\
= & \sum_{Y} P(Y) \log \frac{1}{P(Y)} \\
& - \sum_{X,Y,Z} P(X,Y,Z) \log \frac{1}{P(Y|X,Z)} \\
= & \sum_{X,Y,Z} P(X,Y,Z) \log \frac{1}{P(Y)} \\
& - \sum_{X,Y,Z} P(X,Y,Z) \log \frac{1}{P(Y|X,Z)} \\
= & \sum_{X,Y,Z} P(X,Y,Z) \log \frac{P(Y|X,Z)}{P(Y)} \\
= & \sum_{X,Y,Z} P(X,Y,Z) \log \frac{P(X,Y,Z)}{P(Y)P(X,Z)} \\
= & H(Y) + H(X,Z) - H(X,Y,Z),
\end{aligned}
$$

from which we obtain

$$I(Y;X,Z) = H(Y) + H(X,Z) - H(X,Y,Z). \quad (7)$$

Given that $I(Y;X,Z) \leq 1$ and that $H(Y)$ acts as the baseline reference, it is wise to normalize Eq. (7) as

$$J(Y;X,Z) = \frac{H(Y) + H(X,Z) - H(X,Y,Z)}{H(Y)}. \quad (8)$$

An *index of relevance* is then obtained which evaluates the influence of two variables X, Z with respect to a class variable Y. It takes values between zero (no relevance) and one (maximum relevance). This index acts as the objective function J to be optimized. The reward of using this objective function by a TAFS-like algorithm consists in the possibility of testing it in highly complex domains such as microarray data sets. We name the combination of eTAFS and the objective function in Eq. (8) as the μ-TAFS algorithm.

3.3 Incremental Multivariate Joint Entropy

For a pair of discrete random variables X, Y, it is known that the joint entropy obeys

$$H(X,Y) \geq H(X). \quad (9)$$

This property says that joint entropy is always at least equal to the entropies of the original system: adding a new variable can never reduce the available uncertainty. If we rewrite (9) as an equation

$$H(X,Y) = H(X) + \triangle_X(Y), \quad (10)$$

then $\triangle_X(Y) \geq 0$ represents the *increment* in entropy due to the addition of the variable Y to the system. In a feature selection setting, given Z a class variable, $\tau \subset \mathcal{X}$ the current subset and $H(\tau)$ its joint entropy, if a new feature $X_i \in \mathcal{X} \setminus \tau$ is considered for possible inclusion in the current subset then:

$$H(Z, \tau \cup \{X_i\}) = H(Z,\tau) + \triangle_{Z,\tau}(X_i). \quad (11)$$

It turns out that, to obtain the next calculation, it is computationally far more advantageous to store $H(Z,\tau)$ and calculate the quantity $\triangle_{Z,\tau}(X_i)$ than to compute the full joint entropy $H(Z, \tau \cup \{X_i\})$ directly. In order to obtain this value, an incremental procedure to calculate multivariate joint entropy has been developed as described in what follows.

The incremental multivariate joint entropy (11) must be computed at every evaluation step involving a possible candidate feature X_i to be included

Table 1. *Marginal Entropy Scheme* (MES) tables for one variable (left) and the addition of a second variable (right). $P(\cdot)$ is the probability mass function, obtained from the data (all entropies are in bits)

X_1	$P(X_1)$	$-P(X_1)\ \log P(X_1)$
0	0.538	0.481
1	0.462	0.515
	$H(X_1) =$	0.996

X_1	X_2	$P(X_1, X_2)$	$-P(X_1, X_2)\ \log P(X_1, X_2)$
0	0	0.231	0.488
0	1	0.308	0.523
		$H(X_1, X_2) =$	1.011
1	0	0.154	0.415
1	1	0.308	0.523
		$H(X_1, X_2) =$	0.939

in the current subset τ. Throughout the process, τ is associated with its current *Marginal Entropy Scheme* (MES), a table storing the unique values contained in the data set for its forming features and its corresponding entropy value. An example of a MES table for two binary variables $\{X_1, X_2\}$ is shown in Table 1.

At the initial step ($\tau = \emptyset$) the MES table for the addition of X_1 to $\emptyset$ is indicated in the left part of Table 1. The two unique values and their entropies $H(X_1 = 0) = 0.481$ and $H(X_1 = 1) = 0.515$ are calculated. Let us suppose that a feature X_2 is to be evaluated w.r.t the current subset $\tau = \{X_1\}$. The MES table with its unique forming patterns is indicated in the right part of Table 1. We can see that by introducing X_2 to the current subset τ, four *partitions* are generated for each unique value of X_1: $\{00, 01, 10, 11\}$. In the particular case of $X_1 = 0$, a change in its entropy contribution is produced by the action of X_2 by splitting it into two entropy values: $H(X_1 = 0, X_2 = 0) = 0.488$ and $H(X_1 = 0, X_2 = 1) = 0.523$, for a total entropy of $H(X_1 = 0, X_2) = 1.011$. The increment in entropy $\triangle_\tau$ is obtained as the difference between the current MES (considering the addition of X_2) and the previous scheme (without it), see Table 2.

Table 2. $\triangle_\tau$ computations from the *Marginal Entropy Scheme*, see Table 1

$\triangle_\tau$	$H(X_1, X_2)$	$-P(X_1)\ \log P(X_1)$	difference
$\triangle_\tau(X_1 = 0)$	1.011	0.481	0.531
$\triangle_\tau(X_1 = 1)$	0.939	0.515	0.424
		$\triangle_\tau$	0.954

Finally, this last value is applied to Eq. (11) to obtain the joint entropy $H(X_1, X_2) = H(X_1) + \triangle_\tau(X_2) = 0.996 + 0.954 = 1.950$. The listings in **Algorithms 7** and **8** show the pseudo-code to compute the procedure explained above. The notation $D|\tau$ stands for the restriction of the dataset D to the features in τ.

Algorithm 7: Incremental Multivariate Joint Entropy

```
input  : $\tau$: Current subset;
         $X_i$: Feature to be added;
         $H_\tau$ : Current subset joint entropy;
         $E_\tau$ : Marginal entropies scheme of $H_\tau$;
         $D$ : Data set;
output: $\tau,\ H_\tau,\ E_\tau$
1  if $|\tau| = 0$ then
2      $\tau \leftarrow \{X_i\}$
3      $D \leftarrow$ Sort($D$)
4      $H_\tau \leftarrow$ Joint Entropy of $D$
5      $E_\tau \leftarrow MarginalEntropyScheme(D|\tau)$
6  else
7      $\tau^+ \leftarrow \tau \cup \{X_i\}$
8      $Sort(D|\tau^+)$
9      $E_{\tau^+} \leftarrow MarginalEntropyScheme(D|\tau^+)$
10     $E_{\tau^-} \leftarrow \sum_j E_\tau^j$  //$j$ runs through the values of $\tau$
11     $\triangle_\tau \leftarrow \sum_i E_{\tau^+}^i - E_{\tau^-}^i$
12     $\tau \leftarrow \tau_+$
13     $H_\tau \leftarrow H_\tau + \triangle_\tau$
14     $E_\tau \leftarrow E_{\tau^+}$  // new MES
```

Initial entropy is evaluated in lines 2-5. This first step calculates the starting joint entropy as well as its first MES (lines 4-5), which will be taken as input to the next computation. Note that these two lines can be efficiently implemented as one function using only one loop-cycle with complexity $\theta(|D|)$, where $|D|$ is the number of training instances.

In the **else** part of the **if** clause, the MES is calculated with the addition of X_i to the current subset τ (named E_{τ^+}). Taking into account that previous MES inherits the ordering sequence derived from a previous stage (because of lines 5 and 9), entropies generated by changes in the MES given by $\tau \cup \{X_i\}$ are summed (E_{τ^-}) in groups (line 11) by the newly formed patterns, rendering a one-to-one correspondence between the previous MES and the current MES.

Algorithm 8: *MarginalEntropyScheme* Function

```
input  : $D$ : Data set;
output: $E$
foreach unique value $v$ in $D$ do
    $\Upsilon[v] \leftarrow$ fraction of instances in $D$ with value $v$
$E \leftarrow H(\Upsilon)$  //calculate entropy of this distribution
```

Thus, the entropy contribution $\triangle_\tau(X_i)$, showing the effect of adding X_i to τ, is computed by the difference in both MESs (line 11), being finally added to the current entropy H_τ (line 13). The implementation of lines 10-11 follows the same consideration as that of lines 4-5, and hence complexity is of the same order.

4 Experimental Work

4.1 Data Sets

Five public-domain microarray gene expression data sets are used to test and validate the approach proposed in this work:

- *Colon Tumor*: 62 observations of colon tissue, of which 40 are tumorous and 22 normal, $2,000$ genes [2].
- *Leukemia*: 72 bone marrow observations and $7,129$ probes: 6,817 human genes and 312 control genes [22]. The goal is to tell acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL).
- *Lung Cancer*: distinction between malignant pleural mesothelioma and adenocarcinoma of lung [25]; 181 observations with $12,533$ genes.
- *Prostate Cancer*: used in [49] to analyze differences in pathological features of prostate cancer and to identify genes that might anticipate its clinical behavior; 136 observations and $12,600$ genes.
- *Breast Cancer*: 97 patients with primary invasive breast carcinoma; $24,481$ genes have to be analyzed [52].

To compute the necessary entropies described in previous sections, a discretization process is needed. This change of representation does not often result in a significant loss of accuracy (sometimes significantly improves it [39], [40]); it also offers reduction in learning time [10]. In this work, the CAIM algorithm was selected for two reasons: it is designed to work with supervised data and does not require the user to define a specific number of intervals [31].

4.2 Settings

Provided that the core nature of the μ-TAFS algorithm resides in its stochasticity, multiple runs can be performed and used to obtain better solutions. The experimental design to test the μ-TAFS algorithm measures performance by carrying out 100 different independent runs. In each run, the algorithm is executed on the corresponding dataset and returns the set of all those feature subsets reaching the best found performance (maximum relevance, in this case). The subset that offers the lowest mutual information (MI) among its elements, i.e., the less redundancy, is taken as the subset delivered in this run.

The μ-TAFS parameters are set as follows: $\epsilon = 0.01, T_0 = 0.1$ and $T_{min} = 0.0001$; these are standard settings and are kept constant for all the problems [23]. The cooling function was chosen to be geometric $\alpha(t) = 0.9\,t$, following recommendations in the literature [41].

Table 3. μ-TAFS running performance. *Time* indicates the running time (in minutes) over the 100 executions; J_{eval} is the number of evaluations of J; size is the average size of the final solutions and its standard error

Data set	Time	J_{eval}	size
Colon Tumor	6.41	503,901	6.93 ± 0.06
Leukemia	6.51	506,489	3.36 ± 0.06
Lung Cancer	7.45	560,972	2.58 ± 0.04
Prostate Cancer	98.74	7,119,800	9.85 ± 0.05
Breast Cancer	136.93	10,943,628	9.62 ± 0.03

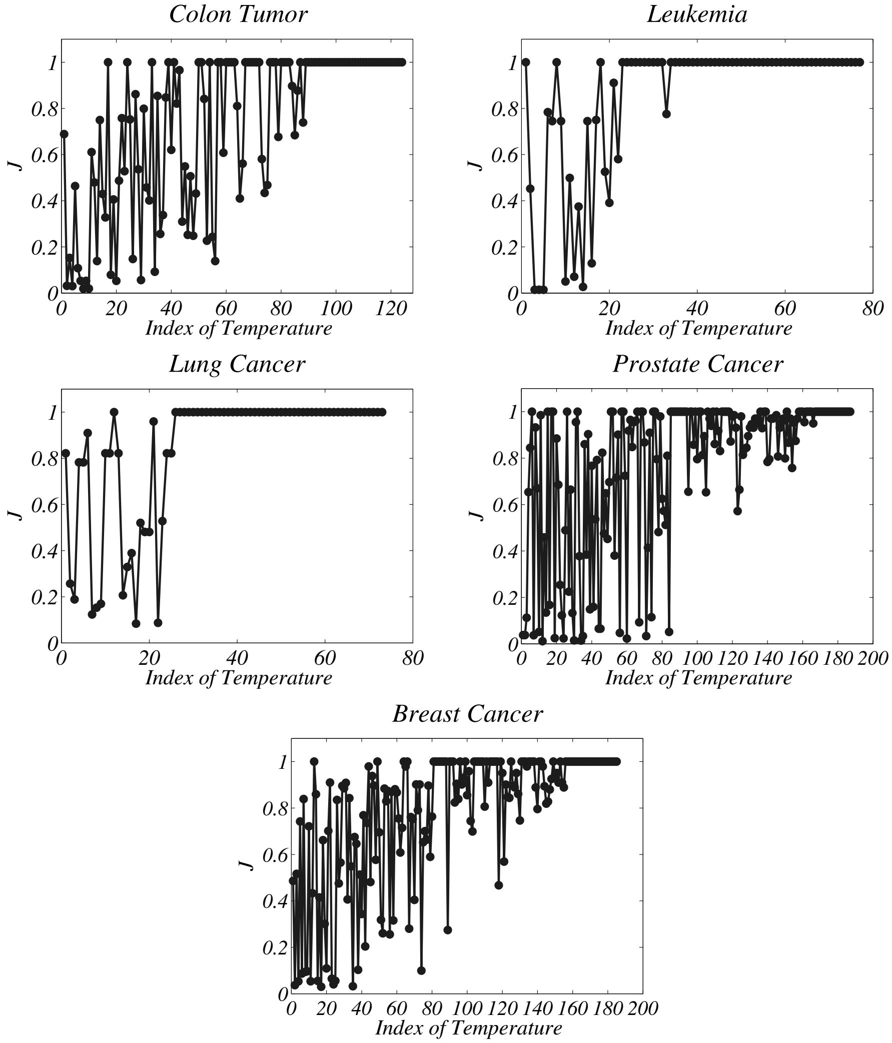


Fig. 4. μ-TAFS search processes. The x-axis is the iteration counter for the outer loop of the algorithm

5 Experimental Results

5.1 μ-TAFS Performance Results

The evolution of μ-TAFS from a high temperature state to a frozen point is depicted in Fig. 4. Highly unstable, i.e., high temperature condition, readings are observed at the initial stages in each of the datasets. As soon as the algorithm becomes more relaxed due to Eq. (1), worse solutions are avoided. The frozen condition is observed at the final stages of each execution, where J values consecutively reach the maximum possible value ($J = 1$) in all cases.

The running performance of μ-TAFS is summarized in Table III. The results show that μ-TAFS yields subsets of considerably low size and also low variability. Notorious readings correspond to *Leukemia* and *Lung Cancer*. It is conjectured that such sizes respond to the nature of the information-theoretic model on discretized data sets, in the sense that only a few genes significantly contribute to increasing the index of relevance given by Eq. (8). On the one hand, working with continuous features, the index would tend to vary smoothly, i.e., generating small increments; as a consequence, more features are added-deleted. On the other hand, discrete features variations are *normalized* by their discretization scheme, so small increments in the real-value are merged into a single discrete value. Therefore, mostly significant increments are truly reflected in its addition-deletion from the current subset.

Computational demands are kept by μ-TAFS considerably low, see Table III. Three of the five problems take 5 to 10 minutes. This is true for the smaller data sets (as *Colon tumor*, with $2,000$ features and *Leukemia*, with just over $7,000$), but also for a larger one, *Lung Cancer*, which needs the processing of more than $12,500$ genes. This behavior is in accordance with the plots in Fig. 4. More complex problems tend to last longer –whereby complexity is related to difficulty in maximizing the objective function, as well as to the dynamics of the forward/backward process. In this sense, *Prostate* and *Breast Cancer* require approximately 1.5 and 2 hours of total processing time. Unfortunately, there is scarcely any reporting on time consumption in recent scientific literature that would enable us to establish a reasonable comparison.

5.2 μ-TAFS Accuracy Results

Seven classifiers were evaluated by means of 10 times 10-fold Cross Validation (10x10 CV), a resampling method that has been suggested as adequate for small sample situations [7]. The chosen classifiers are the k-nearest-neighbors technique (kNN) in which the parameter *k* is the number of neighbors running from 1 to 15, the *Naïve Bayes classifier* (NB), the *Linear* and *Quadratic Discriminant Analysis* classifiers (LDA/QDA), *Logistic Regression* (LR), the *Support Vector Machine* with linear kernel (lSVM) (*regularization parameter* $C = 2^k$, k running from -7 to 7) and the *Support Vector Machine* with radial basis function kernel (rSVM) (*C* parameter in the same conditions, and smoothing parameter $\gamma = 2^k$, k running from -7 to 7)[3]. The non parametric Wilcoxon signed-rank test[4] is used for the (null) hypothesis that the median of the differences between the errors of the best classifier per data set and another classifier's error is zero. The non-parametric Wilcoxon signed-rank test will be used for the (null) hypothesis that the median of differences between classifiers accuracies are zero, at the 95% level of significance.

Table 4. μTAFS: 10x10 mean cross-validation accuracy (*10x10 CV*) complemented with its standard error for the best model in each data set. The *Classifier* column indicates the best method along with best parameters

Data set	Classifier	10x10 CV	size
Colon Tumor	lSVM ($C = 2^1$)	89.19±0.38	5
Leukemia	lSVM ($C = 2^{-7}$)	99.62±0.27	3
Lung Cancer	LR	99.89±0.07	4
Prostate Cancer	kNN ($k = 6$)	95.66±0.21	7
Breast Cancer	rSVM ($C = 2^3, \gamma = 2^{-1}$)	86.90±0.48	6

The obtained solutions are displayed in Table 4. The first fact to note is that the developed algorithm tends to obtain high accuracies that are both very stable and low-sized. This is a very remarkable result, given the big differences among the problems and among the inducers. In particular, *Lung Cancer*, *Leukemia* and *Prostate Cancer* reach remarkably high accuracies, while *Colon Tumor* and specially *Breast Cancer* show lower 10x10 CV readings. In all cases, the subset that delivers this performance is considerably small, having 7 genes or less (and only 3 genes in the *Leukemia* data set). Moreover, all Wilcoxon test p-values signal significant differences ($p < 0.05$) between the best method and all the other methods in the corresponding data set, except for the lSVM vs. LR in *Colon Tumor* ($p = 0.312$).

The results diverge for different classifiers, as it may be reasonably expected. Also, it is very important to assess whether an improvement is consistent or is limited to a certain type of method. In this sense, kNN seems to be the best method for *Prostate Cancer*, LR for *Lung Cancer* and the SVM

[3] For the experiments, we use a MATLAB implementation; specifically, for the SVMs we use the MATLAB interface to LIBSVM [12]. All tests are run on on a regular x86 workstation.

[4] The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test for the analysis of two related samples, or repeated measurements on a single sample. It can be used as an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed. It should therefore be used whenever the distributional assumptions that underlie the t-test cannot be satisfied.

for the other three. The SVM also tends to deliver smaller gene subsets. Given that the SVM parameters were not optimized beyond educated guesses, we think there is room for further improvement in the modeling, specially on the accuracy side.

5.3 Comparison with Other Methods

It is a common practice to compare to similar works in the literature. Unfortunately, the methodological steps are in general very different, especially with respect to resampling techniques, making an accurate comparison a delicate undertaking. Eight references which are illustrative of recent relevant work are analysed and presented in Table 5 (including previous work from the authors). Three of them are filter methods, three other are wrappers and the remaining two are a combination of filters and wrappers. In this table the resampling method, the best classifier and the best result are detailed (the final reported accuracy and number of genes).

The *Colon Tumor* data set presents difficulties in classification, never reaching 90%. The solution delivered by μ-TAFS is comparable with the best known (that of BGS^3 [24]); however, it uses 5 genes against the 9 used by BGS^3. The other difficult problem seems to be *Breast Cancer*. In this data set, μ-TAFS gives the best result among the references consulted, using also less genes and in front of solutions that employ a pure wrapper strategy. For the other three problems, μ-TAFS is also able to yield better solutions compared to other approaches, many of them using a much bigger gene subset.

5.4 Discriminatory Visualization of Selected Features

The genes corresponding to the solutions displayed in Table IV are detailed in Table VI. Moreover, expression levels for each model in the five data sets are given in Fig. 5. It is seen that each model contains genes that are visually identified as presenting abnormal expression levels: *Colon Tumor* genes M76378 and T51288; *Leukemia* genes AFFX-CreX-5_at and L09209; *Lung Cancer* gene 37157_at; *Prostate Cancer* genes 38322_at and 37639_at; and *Breast Cancer* genes Contig14882_RC, Contig53822_RC and Contig57657_RC.

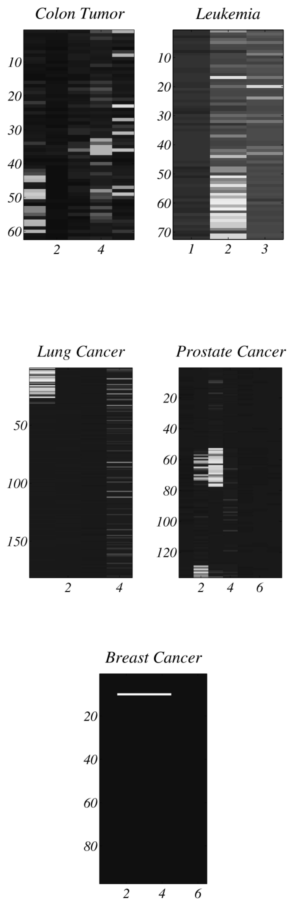


Fig. 5. Expression levels formed as indicated in Table VI. Samples for each data set are distributed as follows: *Colon Tumor*: Tumor 1-41, Normal 42-62; *Leukemia*: Tumor 1-48, Normal 49-72; *Lung Cancer*: Tumor 1-31, Normal 32-181; *Prostate Cancer*: Tumor 1-78, Normal 79-136; and *Breast Cancer*: Tumor 1-46, Normal 47-97

If classification accuracy on the basis of the available data were the only relevant outcome of a modelling method, then feature selection would become a redundant process. Indeed, the *interpretability* of the results is a compulsory requirement in this problem. In a medical context, data visualization in a low-dimensional representation space may become extremely important, helping

http://dx.doi.org/10.13053/CyS-18-2-2014-032

Table 5. Results reported in the literature for the explored problems: (**F**) indicates that the referenced work uses a Filter-Based Algorithm, (**W**) for wrapper and (**FW**) for a combination of both schemes; in parentheses, the size of the subset (number of genes) and the inducer optimized (see text). The $-$ sign indicates that the problem was not studied in the reference. The resampling methods are 10CV (10-fold Cross Validation), 10x10 CV (10 times 10-fold Cross Validation), N-RS (N times Random Subsampling), and 200-B.632 (0.632 bootstrap of size 200)

Work	Validation	Colon Tumor	Leukemia	Lung Cancer	Prostate Cancer	Breast Cancer
[24](**F**)	10x10CV	89.36	97.89	98.84	93.43	83.37
		(9,3NN)	(2,NB)	(4,LR)	(3,10NN)	(12,lSVM)
[9](**F**)	200-B.632	88.75	98.2	–	–	–
		(14,lSVM)	(23,lSVM)	–	–	–
[43](**W**)	10x10CV	85.48	93.40	–	–	–
		(3,NB)	(2,NB)	–	–	–
[53](**W**)	100-RS	87.31	–	72.20	–	–
		(94,SVM)	–	(23,SVM)	–	–
[8](**W**)	50-RS	77.00	96.00	99.00	93.00	79.00
		(33,rSVM)	(30,rSVM)	(38,rSVM)	(47,rSVM)	(46,rSVM)
[27](**FW**)	10x10CV	–	–	99.40	96.30	–
		–	–	(135,5NN)	(79,5NN)	–
[26](**F**)	10CV	–	98.6	99.45	91.18	68.04
		–	(2,SVM)	(5,SVM)	(6,SVM)	(8,SVM)
[45](**FW**)	1-RS	–	–	98.66	67.65	–
		–	–	(8,SVM)	(22,SVM)	–

Table 6. Identification of genes for each model

Data set	Gene ID
Colon Tumor	M76378, H08393, T51849, M19311, T51288
Leukemia	AFFX-CreX-5_at, L09209, X75755
Lung Cancer	37157_at, 33221_at, 107_at, 40790_at
Prostate Cancer	1230_g_at, 38322_at, 37639_at, 32909_at, 660_at 35998_at, 34107_at
Breast Cancer	AB014543, Contig14882_RC, Contig53822_RC, Contig57657_RC, Contig53713_RC, NM_006191

oncologists to gain insights into what undoubtedly is a complex domain. In this work we use a method based on the decomposition of the scatter matrix, with the remarkable property of maximizing the separation between the projections of compact groups of data [33]. Such visualization is illustrated by the plots in Fig. 6. These are scatter plots of 2-D projections of the classes (using the first two eigenvectors of the scatter matrices).

5.5 Biological Evidence

In this section, known biological evidence is presented about the effect of the found gene expressions in each cancer disease. This evidence is assembled by examining recent relevant medical literature.

Colon Tumor

— **M76378** *CSRP1-Cysteine and glycine-rich protein 1*. This gene encodes a member of the cysteine-rich protein (CSRP) family. It may be involved in regulatory processes important for development and cellular differentiation. Hypomethylation, a decrease in the epigenetic methylation of cytosine and adenosine residues in DNA, of CSRIP1 and other genes were confirmed in cancer cells by bisulfite sequencing [54].

— **H08393** *COL11A2-collagen, type XI, alpha 2 (Homo sapiens)*. Two alpha chains of type

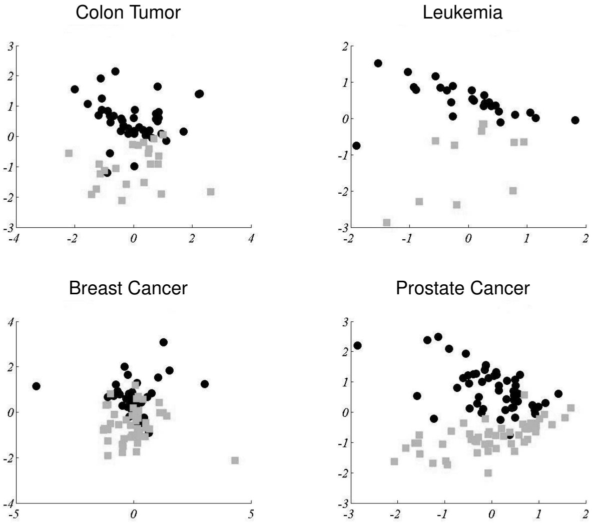


Fig. 6. Visualization of the solutions using the first two eigenvectors of each scatter matrix. Legend: In *Colon Tumor* and *Prostate Cancer* circles represent tumorous samples and squares indicates normal tissue; in *Leukemia* circles indicate ALL cells and squares AML cells; in *Lung Cancer* circles are malignant pleural mesothelioma and squares are adenocarcinoma; in *Breast Cancer* circles indicate relapse samples and squares, non-relapse samples

XI collagen, a minor fibrillar collagen are encoded by this gene [38]. Up-regulation of this gene in the mucosa stromal cells of both familial adenomatosis polyposis and sporadic colorectal cancer has been detected [6].

— **T51849** *EPHB1-Tyrosine-protein kinase receptor elk precursor*. EphB1 is a member of receptor tyrosine kinases of the EphB subfamily and has been positively identified in the development, progress and prognosis of colorectal cancers [48].

— **M19131** *CALM2-calmodulin 2 (phosphorylase kinase, delta)*. Caml2 plays an important role in intracellular calcium signaling, which regulates a variety of cellular processes, such as cell proliferation and gene transcription [5]. Increased expression levels of this gene were found in anaplastic large cell lymphoma cell lines [42].

— **T51288** *ASS1-argininosuccinate synthase (human)*. Arginine, a semi-essential amino acid in humans, is critical for the growth of human cancers as in primary ovarian, stomach and colorectal cancer, whose expression levels read high values [15].

Leukemia

— **AFFX-CREX-5_AT** NOT IDENTIFIED.

— **L09209** *APLP2-amyloid beta (A4) precursor-like protein 2 (Homo sapiens)*. The function of this gene is not fully understood, but it is conjectured that it may play a role in the regulation of hemostasis [19]. This gene was reported as over-expressed by other scientific literature as in [46].

— **X95735_at** *ZYX-ZYXIN*. It is involved in the spatial control of actin assembly and in the communication between the adhesive membrane and the cell nucleus [20]. This is a gene found in many cancer classification studies [22, 14, 11], and it is highly correlated with acute myelogenous leukemia.

Lung Cancer

— **37957_at** *ATG4-Autophagy related 4 homolog A*. Autophagy is the process by which endogenous proteins and damaged organelles are destroyed intracellularly. Autophagy is postulated to be essential for cell homeostasis and cell remodeling during differentiation, metamorphosis, non-apoptotic cell death, and aging [19]. It is activated during amino-acid deprivation and has been associated with neurodegenerative diseases, cancer, pathogen infections and myopathies [44].

— **33221_at** *PAXIP1-PAX interacting (with transcription-activation domain) protein 1*. Member of the paired box (PAX) gene family, this gene plays a critical role in maintaining genome stability by protecting cells from DNA damage [19, 37]. Analysis of pulmonary adenocarcinomas in experiment GDS1650 in [38] records shows over-expression levels of this gene.

— **40790_at** *BHLHE40-basic helix-loop-helix family, member e40*. This gene encodes a basic helix-loop-helix protein expressed in various tissues, it may be involved in the control of cell differentiation [38]. Experiments suggest that loss of DEC1 expression is an early event in the development of lung cancer [21]

— **107_at** *RAB40A-member RAS oncogene family*. This gene encodes a member of the Rab40 subfamily of Rab small GTP-binding proteins that contains a C-terminal suppressors of cytokine signaling box [19]. No medical evidence was found in literature about its role in cancer.

Prostate Cancer

— **1230_g_at** *MTMR11-myotubularin related protein 11*. In experiments on patients with acute lymphoblastic leukemia and with Burkitt lymphoma, three putative oncogenes or tumor suppressor genes were found, one of them was the MTMR11 [50].

— **38322_at** *PAGE4-P antigen family, member 4 (prostate associated)*. This gene is strongly expressed in prostate cancer; and also expressed in other tissues such as testis, fallopian tube, uterus, placenta; besides, it is expressed in testicular cancer and uterine cancer [19].

— **37639_at** *HPN-Hepsin*. Hepsin is a cell surface serine protease and plays an essential role in cell growth and maintenance of cell morphology; it is highly related with prostate cancer, benign prostatic hyperplasia [19].

— **32909_at** *AQP5-aquaporin 5*. Acting as a water channel protein, Aquaporins are a family of small integral membrane proteins linked to other proteins, whose role is the generation of saliva, tears and pulmonary secretions [19]. Experiments with cases of normal and epithelial ovarian tumor tissues suggest an important role of this gene in the tumorigenesis of the latter, and a possible relationship with the ascites formation of ovarian carcinoma [55].

— **660_at** *CYP24A1-cytochrome P450, family 24, subfamily A, polypeptide 1*. This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids [19]. This gene has been reported as responsible for degradation of the active vitamin D metabolite 1,25-dihydroxyvitamin D3 which is known to be antimitotic in prostate cancer cells [17].

— **35998_at** *Hypothetical protein LOC284244 (LOC284244)*. No evidence found.

— **34107_at** *PFKFB2-6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2*. The protein encoded by this gene is involved in the synthesis and degradation of fructose-2,6-bisphosphate, a regulatory molecule that controls glycolysis in eukaryotes [19]. It has been suggested that the induction of *de novo* lipid synthesis –a process that protects cancer cells from free radicals and chemotherapeutics– by androgen requires the up-regulation of HK2 and PFKFB2 [29].

Breast Cancer

— **AB014543** *CLUAP1-clusterin associated protein 1 (Homo sapiens)*. This gene is highly expressed in osteosarcoma, ovarian, colon, and lung cancers [28].

— **Contig57657_RC** *PAK1-p21 protein (Cdc42/Rac)-activated kinase 1 (Homo sapiens)*. This gene encodes a family member of serine/threonine p21-activating kinases, known as PAK proteins, whose role is the regulation of cell motility and morphology [38]. Pak1 is directly related with the Etk/Bmx protein, the latter acts as a control to the proliferation and tumorigenic growth of mammary epithelial cancer cells [3].

— **NM_006191** *PA2G4-Proliferation-associated 2G4, 38kDa (PA2G4)*. Also known as EBP1, this gene encodes an RNA-binding protein involved in growth regulation [19]. The EBP1 has been shown to be a transcriptional corepressor that inhibits the growth of human breast cancer cell lines [1].

— **Contig14882_RC**, **Contig53822_RC**, **Contig53713_RC** NOT IDENTIFIED.

6 Conclusions

An algorithm for feature selection using Simulated Annealing guided by the discrete multivariate joint entropy has been introduced and evaluated. Our experimental results are concerned with the search for small subsets of highly relevant genes in five public-domain Microarray Gene Expression data samples. The excellent results indicate that the algorithm offers a promising general framework for feature selection in very high dimensional data sets.

We have also shown how feature selection appears to be a viable avenue for dimensionality reduction in this field: a reduction of several orders of magnitude in the number of features leads to substantial improvements. This behavior is important, both for computational and scientific reasons. Even without optimization of free parameters (a necessary step in normal conditions), cross-validated wrapper computations with hundreds of thousands of features may take several days of computing time on a standard desk machine. Scientifically, coping with hundreds of features and pretending interpretability of the role of every feature in the model is out of the question in many cases. This is aggravated in the present situation of data scarcity.

The entropic relevance measure has shown to be a good candidate as the objective function to be optimized by the algorithm. The reported classification results are competitive to current standards in analyzing microarray gene expression data with a very affordable execution time. This last aspect should not be overlooked, since database size is constantly growing and the complexity of optimization scenarios (which make extensive use of resampling methods) is ever greater.

One should bear in mind that the excellent reported results do not –by themselves– entail a medical solution to the diseases, a situation that is faced by all statistical and ML solutions. On the contrary, a main goal of exploratory studies of this kind should be directed towards understanding how the variables selected by the model fit in relation to prior knowledge from the medical domain. In this sense, it is our hope that this and related investigations boost studies that unveil the real significance of the findings and advance toward a better understanding of the involved processes.

Acknowledgements

Authors wish to thank Spanish CICyT Project No. CGL2004-04702-C02-02, CONACyT and UABC for supporting this research from its beginning.

References

1. **Akinmade, D., Talukder, A., Zhang, Y., Luo, W., Kumar, R., & Hamburger, A.** (**2008**). Phosphorylation of the erbb3 binding protein ebp1 by p21-activated kinase 1 in breast cancer cells. *British Journal of Cancer*, 98, 1132–1140.

2. **Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., & Levine, A.** (**1999**). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of The National Academy of Sciences USA*, volume 96, IEEE, pp. 6745–6750.

3. **Bagheri-Yarmand, R., Mandal, M., Taludker, A., Wang, R., Vadlamudi, R., Kung, H., & Kumar, R.** (**2001**). Etk/bmx tyrosine kinase activates pak1 and regulates tumorigenicity of breast cancer cells. *Journal of Biological Chemistry*, 276(31), 29403–29409.

4. **Bell, D. & Wang, H.** (**2000**). A formalism for relevance and its application in feature subset selection. *Machine Learning*, 41(2), 175–195.

5. **Bhattacharya, S., Bunick, C., & Chazin, W.** (**2004**). Target selectivity in ef-hand calcium binding proteins. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1742(1-3), 69–79.

6. **Bowen, K., Reimers, A., Luman, S., Kronz, J., Fyffe, W., & Oxford, J.** (**2008**). Immunohistochemical localization of collagen type xi a1 and a2 chains in human colon tissue. *Journal of Histochemistry and Cytochemistry*, 56(3), 275–283.

7. **Braga-Neto, U. & Dougherty, E. R.** (**2003**). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3), 374–380.

8. **Bu, H., Li, G., & Zeng, X.** (**2007**). Reducing error of tumor classification by using dimension reduction with feature selection. *The First International Symposium on Optimization and Systems Biology (OSB 2007)*, pp. 232–241.

9. **Cai, R., Hao, Z., Yang, X., & Wen, W.** (**2009**). An efficient gene selection algorithm based on mutual information. *Neurocomputing*, 72, 991–999.

10. **Catlett, J.** (**1991**). On changing continuous attributes into ordered discrete attributes. *Proceedings of the European working session on learning on Machine learning*, Springer-Verlag New York, Inc., New York, NY, USA, pp. 164–178.

11. **Chakraborty, S.** (**2009**). Simultaneous cancer classification and gene selection with bayesian nearest neighbor method: An integrated approach. *Computational Statistics and Data Analysis*, 53(4), 1462–1474.

12. **Chang, C. & Lin, C.** (**2002**). Libsvm : a library for support vector machines. In http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

13. **Chu, F. & Wang, L.** (**2005**). Applications of support vector machines to cancer classification with microarray data. *International Journal of Neural Systems*, 15(6), 475–484.

14. **Chu, W., Ghahramani, Z., Falciani, F., & Wild, D.** (**2005**). Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21(16), 3385–3393.

15. **Delage, B., Fennell, D., Nicholson, L., McNeish, I., Lemoine, N., Crook, T., & Szlosarek, P.** (**2010**). Arginine deprivation and argininosuccinate synthetase expression in the treatment of cancer. *International Journal of Cancer*, 126(12), 2762–2772.

16. **Duan, K., Rajapakse, J., Wang, H., & Azuaje, F.** (**2005**). Multiple svm-rfe for gene selection in cancer classification with expression data. *IEEE/ACM Transactions on Nanobioscience*, 4(3), 228–234.

17. **Farhana, H., Wahalab, K., Adlercreutzc, H., & Cross, H.** (**2002**). Isoflavonoids inhibit catabolism of vitamin d in prostate cancer cells. *Journal of Chromatography B*, 777(1-2), 261–268.

18. **Filippone, M., Masulli, F., & Rovetta, S.** (**2006**). Unsupervised gene selection and clustering using simulated annealing. In **Bloch, I., Petrosino, A., & Tettamanzi, A.**, editors, *Fuzzy Logic and Applications*, volume 3849 of *Lecture Notes in Computer Science*. Springer, 229–235.

19. **GenCards** (**2009**). Weizmann Institute of Science. http://www.genecards.org/.

20. **GeneAtlas** (**2007**). Université René Descartes - Paris. In http://www.dsi.univ-paris5.fr/genatlas/.

21. **Giatromanolaki, A., Koukourakis, M., Sivridis, E., Turley, H., Wykoff, C., Gatter, K., & Harris, A.** (**2003**). Dec1 (stra13) protein expression relates to hypoxia- inducible factor 1-alpha and carbonic anhydrase-9 overexpression in non-small cell lung cancer. *The Journal of Pathology*, 200(2), 222–228.

22. **Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., & Lander, E.** (**1999**). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.

23. **González, F. & Belanche, L.** (**2008**). A thermodynamical search algorithm for feature subset selection. In **Ishikawa, M., Doya, K., Miyamoto, H., & Yamakawa, T.**, editors, *Neural Information Processing*, volume 4984 of *Lecture Notes in Computer Science*. Springer, 683–692.

24. **González, F. F. & Belanche, L. A.** (**2011**). Parsimonious selection of useful genes in microarray gene expression data. In **Arabnia, H. R. & Tran, Q.-N.**, editors, *Software Tools and Algorithms for Biological Systems*, volume 696 of *Advances in Experimental Medicine and Biology*. Springer New York, 45–55.

25. **Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D., & Bueno, R.** (**2002**). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62, 4963–4967.

26. **Hewett, R. & Kijsanayothin, F.** (**2008**). Tumor classification ranking from microarray data. *BMC Genomics*, 9(2).

27. **Hong, J. & Cho, S.** (**2008**). Cancer classification with incremental gene selection based on dna microarray data. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE, pp. 70–74.

28. **Ishikura, H., Ikeda, H., Abe, H., Ohkuri, T., Hiraga, H., Isu, K., Tsukahara, T., Sato, N., Kitamura, H., Iwasaki, N., Takeda, N., & Nishimura, A. M. T.** (**2011**). Identification of cluap1 as a human osteosarcoma tumor-associated antigen recognized by the humoral immune system. *International Journal of Oncology*, 30(2), 225–233.

29. **Jong-Seok, M., Won-Ji, J., Jin-Hye, K., Hyo-Jeong, K., Mi-Jin, Y., Jae-Woo, K., Park, P. S. W., & Kyung-Sup, K.** (**2011**). Androgen stimulates glycolysis for de novo lipid synthesis by increasing the activities of hexokinase 2 and 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase 2 in prostate cancer cells. *Biochemical Journal*, 433, 225–233.

30. **Kirkpatrick, S.** (**1984**). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34.

31. **Kurgan, L. & Cios, K.** (**2004**). Caim discretization algorithm. *IEEE Trans. on Knowledge and Data Engineering*, 16(2), 145–153.

32. **Li, Y. & Liu, Y.** (**2008**). A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data. *Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB '08. IEEE Symposium on*, pp. 195–200.

33. **Lisboa, P., Ellis, I., Green, A., Ambrogi, F., & Dias, M.** (**2008**). Cluster based visualisation with scatter matrices. *Pattern Recognition Letters*, 29(13), 1814–1823.

34. **Lu, Y. & Han, J.** (**2003**). Cancer classification using gene expression data. *Information Systems*, 28, 243–268.

35. **Meiri, R. & Zahavi, J.** (**2006**). Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3), 842–858.

36. **Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E.** (**1953**). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21.

37. **Munoz, I. & Rouse, J.** (**2009**). Control of histone methylation and genome stability by ptip. *EMBO reports*, 10.

38. **NCBI** (**2007**). National Center of Biotechnology Information. In http://www.ncbi.nlm.nih.gov/.

39. **Ng, M. & Chan, L.** (**2005**). Informative gene discovery for cancer classification from microarray expression data. *IEEE Workshop on Machine Learning for Signal Processing*, IEEE, pp. 393–398.

40. **Potamias, G., Koumakis, L., & Moustakis, V.** (**2004**). Gene selection via discretized gene-expression profiles and greedy feature-elimination. *SETN*, pp. 256–266.

41. **Reeves, C. R.** (**1995**). *Modern Heuristic Techniques for Combinatorial Problems*. McGraw Hill.

42. **Renata, R., Visser, L., der Leij, J. V., Harms, G., Blokzijl, T., Deloulme, J., van der Vlies, P., Kamps, W., Kok, K., Lim, M., Poppema, S., & van den Berg, A.** (**2005**). High expression of calcium-binding proteins, s100a10, s100a11 and calm2 in anaplastic large cell lymphoma. *British Journal of Haematology*, 131(5), 596–608.

43. **Ruiz, R., Riquelme, J., & Aguilar, J.** (**2006**). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39, 2383–2392.

44. **Scherz-Shouval, R., Shvets, E., Fass, E., Shorer, H., Gil, L., & Elazar, Z.** (**2007**). Reactive oxygen species are essential for autophagy and specifically regulate the activity of atg4. *The EMBO Journal*, 26, 1749–1760.

45. **Shah, S. & Kusiak, A.** (**2007**). Cancer gene search with data-mining and genetic algorithms. *Comput. Biol. Med.*, 37(2), 251–261.

46. **Shaik, J. & Yeasin, M.** (**2007**). A unified framework for finding differentially expressed genes from microarray experiments. *BMC Bioinformatics*, 8(1).

47. **Shannon, C. E.** (**1948**). A mathematical theory of communication. *The Bell System Technical Journal.*, 27, 379–423.

48. **Sheng, Z., Wang, J., Dong, Y., Ma, H., Zhou, H., Sugimura, H., Lu, G., & Zhou, X.** (**2008**). Ephb1 is underexpressed in poorly differentiated colorectal cancers. *Pathobiology*, 75(5), 274–280.

49. **Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., & Sellers, W.** (**2002**). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.

50. **Starza, R. L., Crescenzi, B., Pierini, V., Romoli, S., Gorello, P., Brandimarte, L., Matteucci, C., Kropp, M., Barba, G., Martelli, M., & Mecucci, C.** (**2007**). A common 93-kb duplicated dna sequence at 1q21.2 in acute lymphoblastic leukemia and burkitt lymphoma. *Cancer Genetics and Cytogenetics*, 175(1), 73–76.

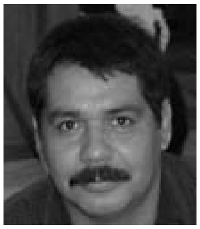
51. **Tang, Y., Zhang, Y., & Huang, Z.** (**2007**). Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3), 365–381.

52. **Vant'Veer, L., Dai, H., Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., & Friend, S.** (**2002**). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 1(415), 530–536.

53. **Wang, L., Zhu, J., & Zou, H.** (**2008**). Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3), 412–419.

54. **Wang, Q., Williamson, M., Bott, S., Brookman-Amissah, N., Freeman, A., Nariculam1, J., Hubank3, M., Ahmed, A., & Masters, J.** (**2007**). Hypomethylation of wnt5a, crip1 and s100p in prostate cancer. *Oncogene*, 26, 6560–6565.

55. **Yang, J., Shi, Y., Cheng, Q., & Deng, L.** (**2006**). Expression and localization of aquaporin-5 in the epithelial ovarian tumors. *Gynecologic Oncology*, 100(2), 294–299.



Félix Fernando González-Navarro holds a Ph.D. in Artificial Intelligence (2011) from the Software Department at the Universitat Politènica de Catalunya (UPC). Currently, he is full-professor at the Universidad Autónoma de Baja California, leading the Artificial Intelligence Lab, where machine learning techniques are applied in the research fields such as Microarray Gene Expression Analysis, Proton Magnetic Resonance Spectroscopy Analysis and Biosensors Modeling. He is reviewer of several international journals and conferences such as MCPR (2012-14), MICAI (11-13), CONIELECOMP(12-14).

Lluís A. Belanche-Muñoz is an associate professor in the Departament de Llenguatges i Sistemes Informtics at the Universitat Politècnica de Catalunya (UPC) in Barcelona, Spain. He received a B.Sc. in Computer Science from the UPC in 1990 and a M.Sc. in Artificial Intelligence in the UPC in 1991. He joined the Computer Science Faculty shortly after, where he completed his doctoral dissertation in 2000. His research involves neural networks and support vector machines for pattern recognition and function approximation, as well as feature selection algorithms, and their collective application to workable artificial learning systems.

Article received on 07/03/2013, accepted on 11/06/2013.