RAIRO-Oper. Res. 50 (2016) 437–449
DOI: 10.1051/ro/2015045

RAIRO Operations Research
www.rairo-ro.org

OPTIMAL DISCRETIZATION AND SELECTION OF FEATURES BY ASSOCIATION RATES OF JOINT DISTRIBUTIONS

Daniele Santoni[1], Emanuel Weitschek[1,2] and Giovanni Felici[1]

Abstract. In this paper we propose a new method to measure the contribution of discretized features for supervised learning and discuss its applications to biological data analysis. We restrict the description and the experiments to the most representative case of discretization in two intervals and of samples belonging to two classes. In order to test the validity of the method, we measured the abundance of different explanatory models that can be derived from a given set of binary features. We compare the performances of our algorithm with those of popular feature selection methods, over three different publicly available gene expression data sets. The results of the comparison are in favour of the proposed method.

Mathematics Subject Classification. 62H30.

Received September 8, 2015. Accepted September 21, 2015.

1. Introduction

In data analysis and data mining one typically wants to find simple models that explain a large amount of observed data, with the purpose of acquiring new knowledge or to understand why certain samples possess certain characteristics. This objective is particularly relevant and challenging when dealing with biological data sets. The realization of effective computational methods that accomplish it is one of the tasks of bioinformatics.

In the analysis of biological data, we often are faced with the following situation: a set of samples of an organism divided into two or more classes, (*e.g.*, tissue from normal and diseased patients), with the related gene expression levels or gene copy number variation, measured on those tissues by a specific technology, is given. In such cases, the selection of small *gene lists* related to the conditions expressed in the samples could be of great value to orient successive experiments conducted by doctors and biologists.

A very general and commonly used paradigm is that of supervised learning, where a model, based on a subset of the features and expressed in a formal and computable language, captures the relevant characteristics of the samples. In this framework, there are several statistical, mathematical, and algorithmic problems. Here, we focus on the discretization, *i.e.*, the transformation of features represented by a continuous measures into

Keywords. Features selection, discretization, data mining.

[1] Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, 00185 Rome, Italy

[2] Department of Engineering, Uninettuno International University, Corso Vittorio Emanuele II, 39, 00186 Rome, Italy.
giovanni.felici@iasi.cnr.it

features represented by a finite number of discrete classes, and the selection of *relevant* subsets of such features from which good models can be constructed.

The topic of discretization and feature selection has drawn the attention of many researchers; several surveys are available to which the non-expert reader may refer for a more detailed insight into the problem (*e.g.*, [6, 8, 17, 22, 23]).

Focusing on more specific contributions, experiments classification *via* microarray data is the object of the work described in [20], with the aim of distinguishing between tumoral and non tumoral cells.

Microarray technology is often used to classify samples based on gene expression. For this reason, many classification methods and tools for microarray experiments analysis are available; among the classification methods, Support Vector Machines (SVM) are often used, due to the continuous nature of the features. In [19] a comparative study of common microarray experiments classification algorithms is performed on seven different cancer types. The authors also show that feature selection and discretization techniques significantly improve the classification rates.

Also in [5] the important role of feature selection emerges: here data on tumor tissues are analyzed with both supervised and non supervised classification techniques, and a feature selection method, based on the discriminating power of the genes, is reported to provide a significant increase in correct classification rates. Microarray classification experiments with SVM are reported in [12, 14, 18], while a simple method, based on nearest neighbour approach, is investigated with success in [30]. In [22] an extensive comparison among microarray classification methods is performed (related to SVMs, nearest neighbour approaches, decision trees, error correcting output codes); here, no method emerges over the others. Also in this study the authors favor the use of feature selection methods to be applied before the classification algorithm.

In the present work a novel and promising approach that integrates discretization and feature selection is presented, with the objective of providing a significant contribution to the step of initial filtering of the features to be used in classification algorithms.

One of the main motivations for this work is that the use of discretized features (namely, binarized features) is well aligned with the concept of genes being over or under-expressed. Moreover, it allows the construction of compact logic models that could more easily represent the information contained in data.

The main idea behind the proposed algorithm is that samples belonging to different classes can be separated by a given feature – completely or partially – if the joint distribution of their expression values, related to that feature, significantly deviates from randomness. Conversely, a joint distribution similar to a distribution of the random sampling of the feature values suggests that the two classes are likely to be not separated by that feature.

We restrict our description and analysis to the simplest case of samples divided into only two classes, referred in the following as class A and class B, and of discretization in two intervals (binarization). Our method can also be applied when both assumptions are removed, even if it leads to slightly more complicated setting. Nevertheless, these assumptions are quite common in bioinformatics applications, where 2 class problems and discretization in 2 intervals (over/under expressed, low/high, *etc.*) are frequent.

The paper is organized as follows. Section 2 describes the ideas behind the computation of the *Z-scores* and the algorithm to compute them (Sect. 2.1). Features are then binarized using class entropy minimization (Sect. 2.2), and sorted in decreasing order of importance according to the methods described in Section 2.3. The biological data sets used for the experiments are briefly described in Section 2.6. The proposed methods based on the integration of *Z-scores* are then compared with some established feature selection methods, described in Section 2.4; the results of the compared methods over the considered data sets are evaluated according to a particular measure that is independent of the classification algorithm adopted (see Sect. 2.5); the results of the experiments are synthesized in Section 3. As pointed out in Section 4, the results reveal the consistency and the validity of our approach.

2. Material and methods

Following the strategy applied in [26], we designed an algorithm to evaluate the *association rate* of any two sets of real values. The main idea behind the algorithm is that, given two sets of real values, if their joint distribution is similar to a random distribution, they are likely to be not associated. On the contrary, a joint distribution that significantly deviates from randomness suggests they are associated.

2.1. Main algorithm – Distance and Z-score

The original algorithm described in [26] was designed to evaluate the association rate of any two letters over a finite alphabet with respect to a given word. In this work we used the same strategy but instead of having occurrences positions of the two letters we have real values divided into two classes. The following description holds for a given feature, but will be later applied to all available features.

Let A and B be two arrays $A = a_1, a_2, \ldots, a_n$ where $a_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ and $B = b_1, b_2, \ldots, b_m$ where $b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, m$ with dimension n and m, respectively.

We define a distance function $D(A, B)$ as follows:

$$D(A, B) : \mathbb{R}^n \times \mathbb{R}^m \Rightarrow \mathbb{R}$$

where

$$D(A, B) = \sum_{i=1}^{n} \text{Min}\{d_i = |a_i - b_j|, j = 1, .., m\}. \tag{2.1}$$

In other words, for each element $a_i \in A$ we look for the closest element $b_j \in B$, and we compute the distance d_i (the absolute value of the difference) between a_i and b_j; then we sum all d_i obtaining $D(A, B)$; this provides a preliminary association rate between the considered sets.

The function $D(A, B)$ is not symmetrical (*i.e.* $D(A, B) \neq D(B, A)$). In addition, the measure $D(A, B)$ is not a valuable parameter to compare different sets in terms of association because it strongly depends on the size of the two arrays and on the peculiarity of their distributions. In order to remove this bias, we compare the value of $D(A, B)$ with the distance between A and a set of $|B|$ random real values (indicated with $\text{Rand}(B)$).

We run this procedure for a given number of iterations (1000 being a reasonably large number), and we verify that this number is large enough to guarantee a satisfactory approximation. To compare $D(A, B)$ with the distribution of $D(A, \text{Rand}(B))$, we compute the *Z-score* $Z(A, B)$ as follows:

$$Z(A, B) : \mathbb{R}^n \times \mathbb{R}^m \Rightarrow \mathbb{R}$$

$$Z(A, B) = \frac{D(A, B) - \overline{D(A, \text{Rand}(B))}}{\sigma_{D(A, \text{Rand}(B))}} \tag{2.2}$$

where $\overline{D(A, \text{Rand}(B))}$ and $\sigma_{D(A, \text{Rand}(B))}$ are the average and the standard deviation of all the computed $D(A, \text{Rand}(B))$.

It is worth recalling that the *Z-score* function is not symmetrical and the two values $Z(A, B)$ and $Z(B, A)$ can assume very different values. A negative $Z(A, B)$ means that all the elements of A are closer to at least an element of B than expected by chance. On the other hand a positive $Z(A, B)$ means that the occurrences of a are typically farther from the occurrences of b than expected by chance. When both $Z(A, B)$ and $Z(B, A)$ are largely positive we can conclude that the two considered arrays are separated classes of objects. On the contrary when both $Z(A, B)$ and $Z(B, A)$ are largely negative we can conclude that the two arrays are indistinguishable from one another.

Z-score measures the deviation from randomness in terms of standard deviations so it can be considered an *absolute* measure that allows to compare values associated to any couple of sets.

We also note that the time complexity of the algorithm for the computation of the *Z-scores* for a given variable is bounded by $n * \log(n)$, where n is the number of samples; this is due to the need of sorting the two classes arrays. The computation of $D(A, B)$ is linear once the two arrays are sorted, and the time complexity for the computation of *Z-score* is also linear with respect to the number of samples, being the number of iterations constant.

2.2. Discretization *via* class entropy minimization

There are many techniques that can be used to discretize a continuous variable into a discrete one with 2 or more intervals. Here we have the purpose of using the discretized variable to distinguish among samples of different classes, a case in which the ideal discretization would be such that each interval contains only elements of a single class. In the data mining terminology, such intervals may be defined *pure* w.r.t. the class of the samples. The purity of a discretization can be measured in several ways, and we adopt here the *Class Entropy* (CE), a widely used measure adopted by many data mining algorithms just for the purpose of discretization (see *e.g.* Decision Trees, [11]). In general, if N_c is the number of classes, N is the number of intervals, $f_{ij}, i = 1, \ldots, N_c, j = 1, \ldots, N$ is the proportion of samples of class i in interval j, and $n_j, j = 1, \ldots, N$ is the number of samples falling in interval j, Class Entropy is defined as

$$CE = -\sum_{j=1}^{N} n_j \times \sum_{i=1}^{N_c} f_{ij} \times \log(f_{ij}). \tag{2.3}$$

It is easy to see that the CE of a discretized feature is equal to 0 when the discretization is perfect, while grows more and more as samples of different classes merge into the intervals of discretization. It is therefore advisable to adopt a discretization where CE reaches its minimum. This can be done quickly when only two intervals are required (binarization): first we order the samples in increasing order of value of the feature; second, each midpoint between two consecutive samples of different classes is considered a potential cutpoint; third, we evaluate the CE associated to the binary variables associated to that cutpoint. Finally, the cutpoint with lower CE is selected as the best cutpoint and a binary feature is created; then we place in the first interval the samples with feature value below the cutpoint, and in the second interval those with feature value above the cutpoint.

This procedure requires simple computations and a number of iterations linearly bounded by the number of samples considered.

2.3. Selection of features *via* $Z(A, B)$ and $Z(B, A)$

We propose to select the promising features according to the values of their $Z(A, B)$ and $Z(B, A)$ values. Such choice is based on the intuition that, although features with very high values of both $Z(A, B)$ and $Z(B, A)$ are supposed to provide close to perfect separation between the A and B subset of samples, they do not tell the whole story of the information contained in the data. *E.g.*, if no such very good symmetric separating feature exists, we can still trust two features to do a good separating job together, if one is large and positive on $Z(A, B)$ and the other one on $Z(B, A)$. We in fact know, straightforwardly from their definition (see Sect. 2.2), that when, for a given feature, $Z(A, B)$ is largely positive and $Z(B, A)$ is not, it necessarily holds that a large portion of samples in A is distinct from the samples in B by that feature, but also that a large portion of the samples in B is mixed up with samples (possibly, few) of A. Such consideration leads to the conclusion that, while features with $Z(A, B)$ and $Z(B, A)$ both large and positive are in general to be preferred, also those with a positive value of only one of the two can serve well for separation, especially when the first type does not abound in the data.

According to this principle we propose to select those features with the highest value of a combination of the two values. In the experiments we consider three basic simple combinations, and namely:

(1) $WSUM$: $Z(A,B)+Z(B,A)$; the sum of the two indicators is used to rank the features in descending order;

(2) $WMAX$: $\max(Z(A,B),Z(B,A))$; the maximum of the value of two indicators is used to rank the features in descending order;

(3) $WPROD = \begin{cases} \max(Z(A,B),Z(B,A)) \; if \;\; \min(Z(A,B),Z(B,A)) \leq 0; \\ \max(Z(A,B),Z(B,A)) \text{ otherwise;} \end{cases}$

the product of the value of two indicators is used to rank the features in descending order – only when both values are positive, else the maximum value is used.

While WSUM and WMAX are derived straightforwardly from the *Z-scores*, WPROD requires a little more manipulation to take into account the uncommon case of negative Z-scores. When this situation arises we use WMAX, in order to avoid that the negative value of one of the scores compromises the possibly positive and large value of the other. Once the features have been ranked according to one of the criteria listed above, the top K features are selected, binarized (as from Sect. 2.2) and tested with the method that will be described in Section 2.5. Varying the value of K in a reasonable range, we can thus analyze the quality of the information provided by features selected by each criterion. The same framework is then adopted for other established methods that are used to rank the features (as described in Sect. 2.4: Fold Change (FC), Information Gain (IG), Gain Ratio (GR), Relief (RLF)).

2.4. Methods used for comparison

As already pointed out, the focus of this paper is on fast and effective methods for discretization and feature selection, to be used for supervised learning in applications where the data sets are characterized by large dimensions, in particular as far as the number of features is concerned. The proposed method is thus a *filter* method and to ascertain its validity we compare it with the most commonly used methods of the same type, keeping in mind that when smaller data sets or large computation times are available a *wrapper* approach may lead to better final performance (for a more detailed discussion of the differences between the two classes of feature selection methods the interested reader may refer to [6]).

We thus consider 4 different methods:

(1) `Fold Change` is one of the most established methods to ascertain the presence of over- and under-expressed genes in bioinformatic applications. It is obtained as the ratio of a measure of the expression for one class of patients over the other(s). It can be then transformed with several techniques and its statistical significance can be evaluated. For an interesting discussion of this measure and of its drawbacks, see [13]. In the experiments conducted in this paper, we use the $\log_2$ of the ratio of the expression computed on the average of the two classes. Over and under-expressed weights are considered equally significant for the purpose of feature selection. In the following, it is referred to as `FC`.

(2) `Relief`. This method evaluates the value of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class, and is designed to operate on both discrete and continuous class data [21]. Features are evaluated through few heuristics that assess the feature relevance using a distance measure. A feature is given larger weight if it can clearly distinguish two instances of different classes w.r.t. two instances of the same class. In the following, it is referred to as `RLF`.

(3) `Information Gain`. In the context of supervised learning, the term is sometimes used synonymously with mutual information. A feature is considered as a random variable, and the Information Gain is based on the ratio between the prior distribution of that variable given the predicted class and its posterior distribution

conditioned to the class observed throughout the data. This concept is used to define a preferred sequence of attributes and narrow down the feature set. Details are provided in [28]. In the following, it is referred to as `IG`.

(4) `Gain Ratio`. Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. It can be considered as a modification of IG designed to deal more effectively with features that take a large number of discrete values. In the following, it is referred to as `GR`.

For `RLF`, `IG`, and `GR` the implementation available in the open source data mining software *WEKA 3.7* (see [29]) has been used, *via* the `ReliefAttributeEval`, `InfoGainAttributeEval` and `GainRatioAttributeEval` methods, paired with the `Ranker` subset evaluator. `FC` has been implemented directly in the software used for the computation of the $Z(A, B)$ and $Z(B, A)$ values.

2.5. Evaluation of subsets of binary features

Consider a data set made of samples of 2 classes (A and B) described by a finite number m of binary features. In this setting, a good subset of features should enable to always separate elements of one class from elements of the other based on the binary values of the features in the subset. In other words, such subset of binary features must obey the following property: for each couple of samples, where one belongs to class A and the other belongs to class B, there must exist at least one binary feature whose value differs in the two samples. It can easily be derived that a subset of features with this property contains at least one subset of the binary features that can be combined into a Disjunctive Normal Form in propositional logic rule (DNF) that holds *true* for elements of A and *false* for elements of B, thus providing a perfect separation of the two sets. The property above is at the basis of the so called *minimum test collection problem* used for feature selection (see [15]) and of other feature selection methods, used for example in [4,7,10].

Here we adopt the same framework to evaluate the quality of a feature set, with the objective of estimating how many different models that have the very general form of a DNF rule can be extracted from a given subset of features, under the assumption that the number of models contained in a feature set represents a good measure of the information there contained to separate the A and B samples [9]. In addition, if two subsets of features F_1 and F_2 have the same dimension, it is to be preferred the one that is guaranteed to contain more models. Extending this conclusion to feature selection methods, we can compare two methods of the same dimension by using the estimate of how many DNF models – at least – they contain.

Thus, consider the following.

Let F be a subset of the features, with $|F| = h$, and let x_{ijk} be equal to 1 if feature $k \in F$ has different values in sample $i \in A$ and sample $j \in B$, and 0 otherwise. Then, define $\alpha_{ij} = \sum_{k \in F} x_{ijk}$, and

$$\alpha_{\min} = \min_{i \in A, j \in B} (\alpha_{ij}). \tag{2.4}$$

it is now easy to see that $\alpha_{\min}$ is the desired lower bound on the number of DNF models contained in F. Given two feature selection methods, say m_1 and m_2, and the subsets of features of dimension h from them derived, we indicate with $\alpha^h_{\min}(m_1)$ and $\alpha^h_{\min}(m_2)$ the associated values of $\alpha_{\min}$ on subsets of features of dimension h.

We can conclude that method m_1 dominates m_2 if

$$\alpha^h_{\min}(m_1) \geq \alpha^h_{\min}(m_2), \forall h \in H \tag{2.5}$$

where H is an interesting set of values for h, typically $1 \leq h \ll m$ (we recall that m is the total number of features).

2.6. Biological data sets

In this Section, we briefly present the analyzed data sets and explain in detail how they have been obtained from microarrays, which are semiconductor devices able to extract gene expression levels with a unique parallel experiment [27]. Three different microarray data sets are used in this study.

- The first one is a large gene expression data set from the anti-NGF (nerve growth factor) AD11 transgenic mouse model [4], related to Alzheimer's disease in mouse brain. As explained in [4], the gene expression profiling was obtained with a standard two-color protocol by Agilent Technologies [2] by using the Agilent scanner G2564B, equipped with two lasers (532 nm and 635 nm). The images were analyzed by Agilent Feature Extraction software and normalized by the Lowess algorithm [25]. The data set is composed of a total of 142 samples and 20 027 features (genes) divided into two classes: 72 classified as cases and 70 as controls [4]. In the report of the experiments this data set will be referred to as `ALZ`.

- The second data set, described in [16], is related to Leukemia and was obtained by using Affymetrix microarrays [1] from acute leukemia patients. RNA was hybridized to high-density oligonucleotide microarrays and the quantitative expression level for human genes was extracted. The data were subjected to *a priori* quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image. A normalization of the expression levels for each gene and sample was performed such that the mean is 0 and the standard deviation is 1. The data set contains a total of 72 samples and 7 129 features (genes) divided into two classes: 47 classified as Acute Lymphoblastic Leukemia (ALL) and 25 as acute myeloid leukemia (AML) [16]. In the report of the experiments this data set will be referred to as `LEU`.

- The third data set is composed of gene expression profiles obtained from Central Nervous System Embryonal Tumour patients [24]. In this study, RNA was extracted from frozen specimens and was analyzed with oligonucleotide microarrays containing probes for 6 817 genes. Arrays were scanned on Affymetrix scanners and the expression value for each gene was calculated using GENECHIP software [2]. Minor differences in microarray intensity were corrected using a linear scaling method [3] and the data were normalized by standardizing each column (sample) to have zero mean and unit variance. The data set is composed of a total of 60 samples divided into two classes: 39 Medulloblastoma Survivors and 21 Treatment Failures [24]. In the report of the experiments this data set will be referred to as `EMB`.

3. Results

In this section we evaluate the ability to rank the features in an interesting order expressed by 3 different combination of the $Z(A,B)$ and $Z(B,A)$ values (namely: the WSUM, WMAX, and WPROD presented in Sect. 2.3), with the ranking provided by 4 well established feature selection methods (namely: the FC, RLF, IG, and GR, presented in Sect. 2.4). The comparison adopts the lower bound of formula $\alpha_{\min}$ (see formula (2.4)) as a direct measure of the quality of the feature selection method. The comparison is conducted on 3 different data sets, that for dimension, variety and context appear to be a good benchmark for evaluating the validity of the proposed method in bioinformatics applications (data sets described in Sect. 2.6).

The first set of results that we show is the disposition of the features (*i.e.*, genes) in the $Z(A,B)$-$Z(B,A)$ plane, for the 3 data sets considered, indicated with ALZ, EMB, LEU, in Figures 1, 2, and 3. In all the 3 cases the features are densely located close to the origin of the plane, having a small value for both the coordinates and resulting therefore non interesting for separation. More interestingly, several features exhibit very large positive values for one of the two coordinates and a small value for the other (points close to the x or to the y positive semi-axes). According to the definition of *Z-score* given in Section 2.1, these features are supposed to do a good job in separating a large portion of one set from a smaller portion of the other, but are not sufficient – alone – to separate precisely all the data. On the other hand, a combination of features that are lying close to the two positive semi-axes and far from the origin would presumably provide a good global information for separation. The planes also provide a clear information on the presence of few features that have a significant (although, not extreme) balanced contribution of the two scores: those are the features that should provide, if taken alone, the best separations. A quick glance at the 3 planes also provides an interesting view point on the different structures of the 3 data sets, ALZ and EMB appearing more clearly separable by few genes (as we

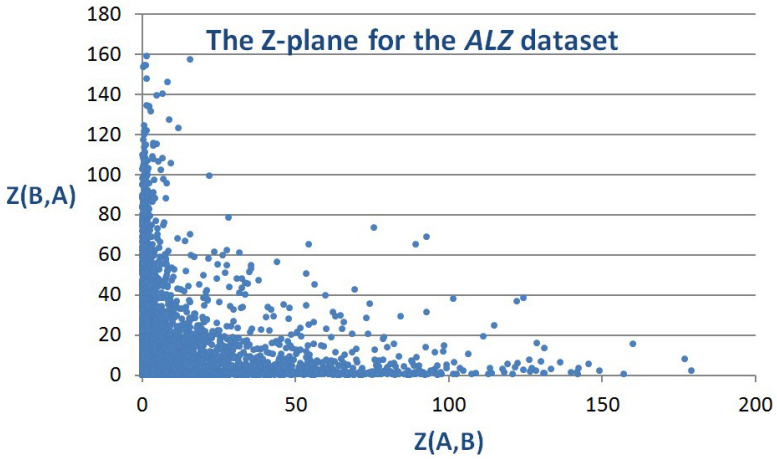

FIGURE 1. The 20 034 features of the ALZ data set represented on the $Z(A,B), Z(B,A)$ plane.

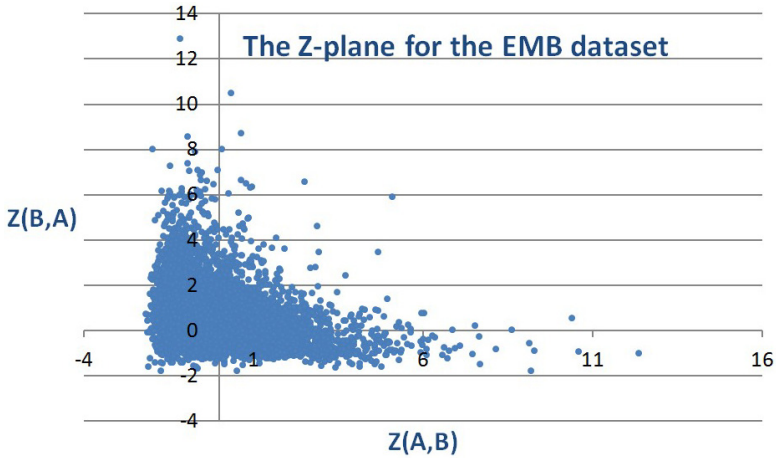

FIGURE 2. The 6 817 features of the EMB data set represented on the $Z(A,B), Z(B,A)$ plane.

have genes that occupy the central portion of the plane) while LEU may require the combination of more genes to obtain similarly good separating models.

It is important to note that the 3 measures adopted (WSUM, WMAX, WPROD) should, in different ways, allow to consider this interpretation of the data, resulting in larger values when at least one of two scores is large; as opposed to WSUM and WPROD, WMAX considers only the larger of the 2 scores, and therefore may assign lower ranks to features that are significant in both scores, but not large enough to compete with other features that have only one very large score.

The role of the combination of the *Z-scores* to identify good features sets is discussed with greater detail in the next section, where the bound on the number of DNF models expressed by formula (2.4) is computed for all methods in relation with the number of selected features.

3.1. *Z-score* as a predictor of differential expression

We turn now to consider the α_{min} value for the different methods. Charts depicted in Figures 4–6 compare the values of this indicator associated with the top portions of the list of genes obtained from the non-decreasing

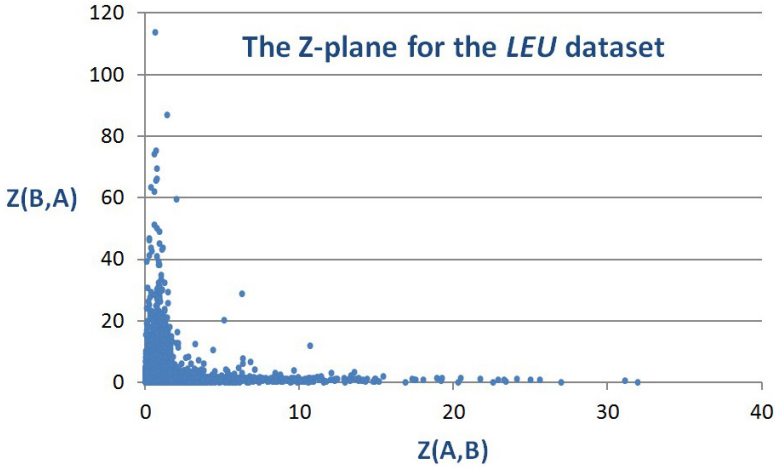

FIGURE 3. The 7 129 features of the LEU data set represented on the $Z(A, B), Z(B, A)$ plane.

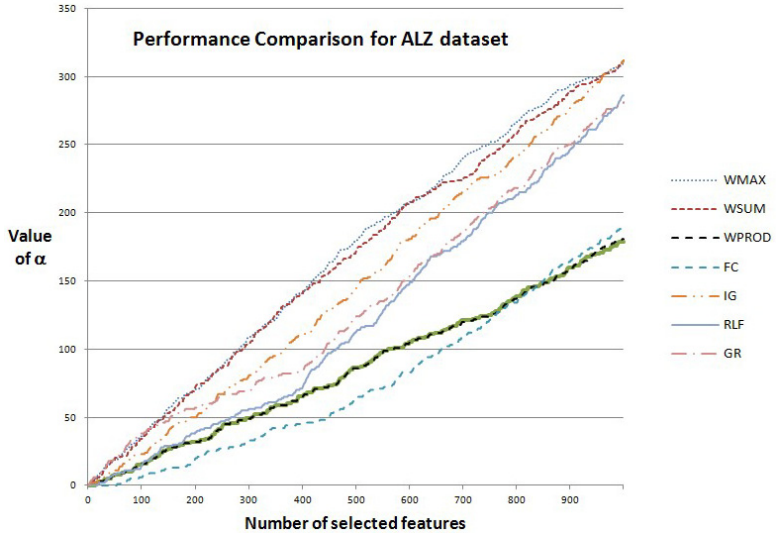

FIGURE 4. Comparison of the methods for the ALZ data set. The chart reports the value of the $\alpha_{\min}$ indicator for the 7 methods considered, for 1 to 1000 selected features.

ranking of the 7 methods. *E.g.*, consider the data set ALZ for the method WMAX. For each feature, we compute the 2 *Z-scores* and the associated WMAX value; then, we rank the features from the largest to the smallest WMAX value and consider all subsets of features composed of the first $1, 2, \ldots, 1000$ features of the list.

Clearly, we expect the value of $\alpha_{\min}$ to increase, or remain the same, as the number of features increases; the steepness of the growth of $\alpha_{\min}$ as the length of the list increases can be directly related with the quality of the selection method that generated that list. Then methods with higher values of $\alpha_{\min}$ should be considered preferable – having once again acknowledged the main assumption of this scheme – two class separation and binary discretization.

Figures 4–6 show with clear evidence that the methods based on *Z-score* perform better than the others for feature sets with dimension ranging from 1 to 1000 in the 3 data sets; the only exception is the behaviour of

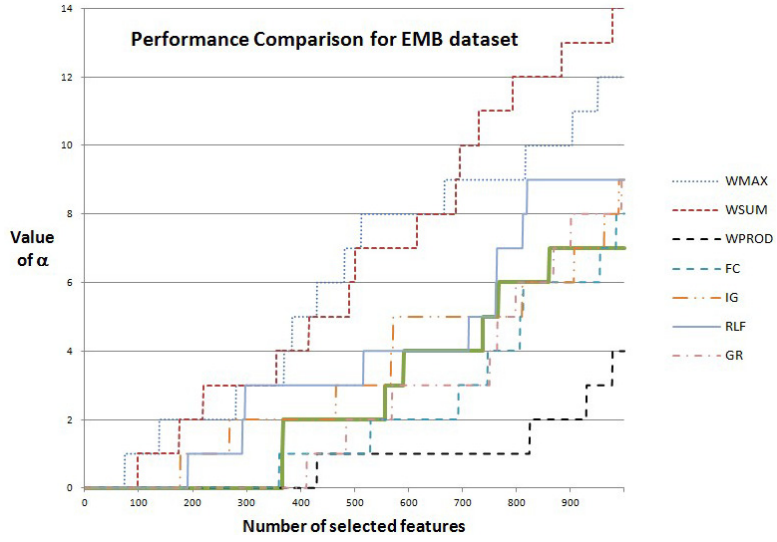

FIGURE 5. Comparison of the methods for the EMB data set.The chart reports the value of the $\alpha_{\min}$ indicator for the 7 methods considered, for 1 to 1000 selected features.

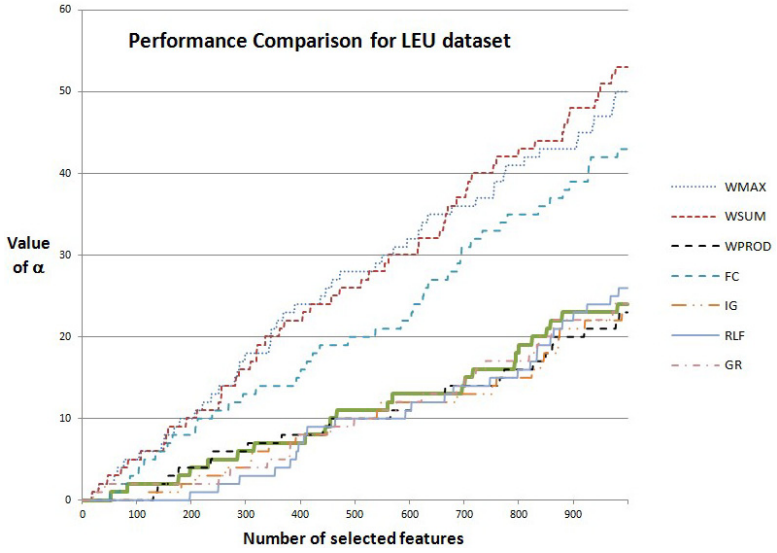

FIGURE 6. Comparison of the methods for the LEU data set.The chart reports the value of the $\alpha_{\min}$ indicator for the 7 methods considered, for 1 to 1000 selected features.

GR in the ALZ experiments, where it exhibits a slightly better behaviour for features sets of small dimension. Apart from this exception, WMAX and WSUM strongly dominate the others, with a marked advantage of the first over the second.

A more detailed evaluation is then provided in Figures 7 and 8. Here we count how many times one method provides the largest value of $\alpha_{\min}$ for a given dimension of the feature set. In presence of ties, we assign the win *ex aequo* and count a "win" for all the methods that reach the maximum value. Such analysis is cumulated over the 3 data sets, thus providing a global indication on the overall performances of the lists based on *Z-scores*.

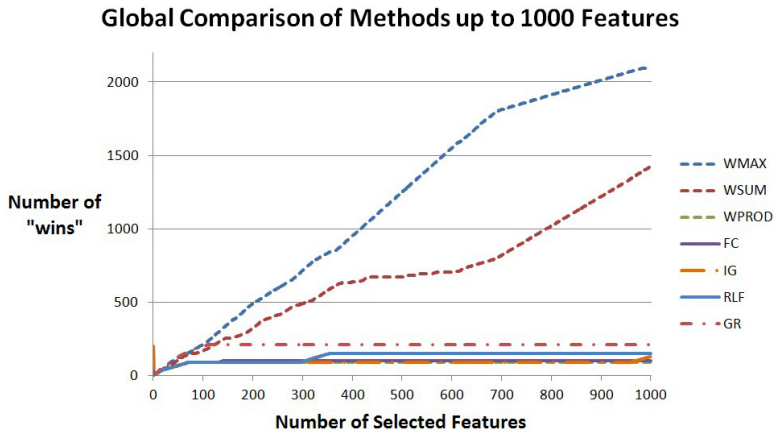

FIGURE 7. Aggregate evaluation of the 7 methods. On the y-axis the number of times a given method provided the best result, for for 1 to 1000 selected features.

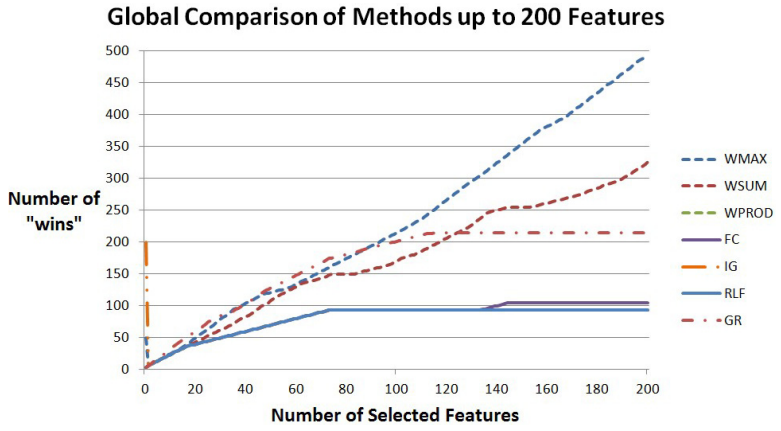

FIGURE 8. Zoom-in aggregate evaluation of the 7 methods. On the y-axis the number of times a given method provided the best result, for for the first 200 selected features.

Results for feature sets with dimension 1–1000 are depicted in Figure 7, where it is clearly shown that, in the long run WMAX outperforms all the others and in particular its direct competitor WSUM. On the other hand, if we zoom in the chart (feature sets with dimension 1–200, Fig. 8) – we again see that GR performs comparably well in the initial ranges – due to its initial supremacy in ALZ data (see Fig. 4); in the initial ranges we can and also see that WSUM is definitely comparable with WMAX. This behaviour of WSUM and WMAX has a convincing interpretation: in the top part of the list WSUM picks the few good isolated features that occupy the central portion of the plane; as these features are in scarce number, WMAX's list starts to dominate for larger feature sets (approx. >60) where the remaining interesting features are those that have high values on only one of the 2 *Z-scores*.

4. Conclusions

In this work we addressed the problem of discretization and selection of features for supervised learning.

We focused on the most representative case of discretization in two intervals – also referred to as *binarization* – and of samples belonging to two classes, although the method could be easily and naturally extended to the

more general case of multiclass data. The problem was faced from a new perspective, using a particular distance and its deviation from randomness as a criterion to select features; in this framework, a good feature is a feature that shows a non-random behaviour.

The interestingness of this measure is that it is multidimensional: *i.e.*, it is represented by a different value for each class of the sample, associated to the degree by which the observed positions of the elements of that class deviate from the distribution of their random positioning, keeping fixed the elements of the other classes. In the simple case of 2-class classification, to which the results here presented are restricted, we obtain 2 values for each features, named their *Z-scores*.

The use of these two measures *together* plays an important role in bringing to evidence the importance of certain features that may else be overlooked by other methods.

We propose a way to measure the quality of feature sets based on the identification of a lower bound on the number of different DNF models that can be expressed with the selected features. Such bound is used to measure the quality of a set of features; the higher the value of the bound, the higher the quality of the set. Using this measure, we have compared different strategies to combine the *Z-scores* of a feature, coming to the conclusion that the WMAX (the maximum of the two scores) and WSUM (the sum of their values) work well, although when the dimension of the selected set is large, WMAX dominates WSUM. Additionally, we compared our feature selection method with other well established filter methods, and show that the lists of features selected on the basis of the *Z-scores* perform better on the considered data sets.

The main result of this paper consists in the introduction of a new feature selection method, that is able to score and rank features for supervised learning in a more effective way than others well established feature rankers. The experimental results highlight the potential advantages of using this method in the analysis of biological data sets, where few relevant continuous features must be extracted from large sets, and properly discretized. Moreover, the proposed method adapts well to data sets coming from different biological contexts, is able to highlight different structural characteristics of the samples, and performs robustly with respect to data dimensions.

Such facts motivate additional research, testing, and developments of the method; in particular, future work will be directed towards the cases with more than 2 classes and towards the effect that such multidimensional *Z-scores* can have on the selection of good features.

References

[1] Affymetrix technologies. www.affymetrix.com.
[2] Agilent technologies. www.genomics.agilent.com.
[3] Affymetrix, Affymetrix Microarray Suite User Guide. Affymetrix, Santa Clara, CA, Version 5 edn. (2001).
[4] I. Arisi *et al.*, Gene expression biomarkers in the brain of a mouse model for alzheimer's disease: mining of microarray data by logic classification and feature selection. *J. Alzheimer's Disease* **24** (2011) 721–738.
[5] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer and Z. Yakhini, Tissue classification with gene expression profiles. *J. Comput. Biol.* **7** (2000) 559–583.
[6] P. Bertolazzi, G. Felici, P. Festa and G. Lancia, Logic classification and feature selection for biomedical data. *Comput. Math. Appl.* **55** (2008) 889–899.
[7] P. Bertolazzi, G. Felici and E. Weitschek, Learning to classify species with barcodes. *BMC Bioinform.* **10** (2009) 1–12.
[8] P. Bertolazzi, G. Felici and G. Lancia, Application of Feature Selection and Classification to Computational Molecular Biology. In *Biological Data Mining*, edited by S. Lonardi and J.K. Chen. Chapman & Hall (2010) 257–294.
[9] P. Bertolazzi, G. Felici, P. Festa, G. Fiscon and E. Weitschek, Integer programming models for feature selection: new extensions and a randomized solution algorithm. *Eur. J. Oper. Res.* **250** (2015) 389–399.
[10] E. Boros, T. Ibaraki and K. Makino, Logical analysis of binary data with missing bits. *Artif. Intell.* **107** (1999) 219–263.
[11] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA (1984).
[12] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M Ares, Jr and D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97** (2000) 262–267.
[13] M.R. Dalman, A. Deeter, G. Nimishakavi and Z.-H. Duan, Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinform.* **13** (2012) 1471–2105.
[14] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinform.* **16** (2000) 906–914.

[15] M.R. Garey and D.S Johnson, Computers and Intractability : A Guide to the Theory of NP-Completeness. *Series Books Math. Sci.* Edited by W.H. Freeman (1979).

[16] T.R. Golub *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** (1999) 531–537.

[17] I. Guyon and A. Elisseeff, An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** (2003) 1157–1182.

[18] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines. *Machine Lear.* **46** (2002) 389–422.

[19] H. Hu, J. Li, A.W. Plank, H. Wang and G. Daggard, A comparative study of classification methods for microarray data analysis. In *AusDM* (2006) 33–37.

[20] T. Jirapech-Umpai and S Aitken, Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinform.* 148 (2005).

[21] I. Kononenko, Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*. Springer (1994) 171–182.

[22] T. Li, C. Zhang and M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinform.* **20** (2004) 2429–2437.

[23] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers (2000).

[24] S.L. Pomeroy *et al.*, Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415** (2002) 436–442.

[25] J. Quackenbush, Microarray data normalization and transformation. *Nature Genet.* **32** (2002) 496–501.

[26] D. Santoni and E. Pourabbas, Automatic detection of words associations in texts based on joint distribution of words occurrences. To appear in *Comput. Intell.* (2015) DOI:10.1111/coin.12065.

[27] M. Schena, D. Shalon, R.W Davis and P.O Brown, Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* **270** (1995) 467–470.

[28] M. Tom, Machine Learning. The Mc-Graw-Hill Companies (1997).

[29] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005).

[30] H. Xiong and X.-W Chen, Kernel-based distance metric learning for microarray data classification. *BMC Bioinform.* **7** (2006) 299.