Weighted Distance Based Discriminant Analysis: The R Package WeDiBaDis

by Itziar Irigoien, Francesc Mestres and Concepcion Arenas

Abstract The **WeDiBaDis** package provides a user friendly environment to perform discriminant analysis (supervised classification). It can be suitable when the user is interested in the problem of constructing a discriminant rule on the basis of distances between a relatively small number of instances or units of known unbalanced-class membership measured on many (possibly thousands) features of any type. This is a current situation when analyzing genetic biomedical data. This discriminant rule can then be used both, as a means of explaining differences among classes, but also in the important task of assigning the class membership for new unlabeled units. Our package implements two discriminant analysis procedures in an R environment: the well-known distance-based discriminant analysis (DB-discriminant) and we introduce a novel classifier rule that we will call weighted-distance-based discriminant (WDB-discriminant). This new procedure is based on an improvement of the DB rule taking into account the statistical depth of the units. This article presents both classifying procedures and describes their implementation in detail. We illustrate the use of the package using an ecological and a genetic experimental examples. Finally, we illustrate the effectiveness of the new proposed procedure (WDB) and compare it with DB. This comparison is carried out using thirty-height high-dimensional class-unbalanced cancer data sets, three of them including clinical features. **WeDiBaDis** is an easy to use package addressed to the biological and medical communities, and in general, to researchers interested in applied studies.

Introduction

Discriminant analysis (supervised classification) is used to differentiate between two or more naturally occurring groups based on a suite of discriminating features. This analysis can be used both as a means of explaining differences among groups, but also to classification, that is, to develop a rule based on features measured on a group of units with known membership (so-called training set), and to use this classification rule to assign the class membership to new unlabeled units. Classification is used by researchers in a wide variety of settings and fields including biological and medical sciences. For example, in biology it is used for taxonomic classification, morphometric analysis for species identification and to study species distribution. Discriminant analysis is applicable to a wide range of ecological problems, as testing for niche separation by sympatric species or for the presence or absence of a particular species. Marine ecologists commonly use discriminant analysis to evaluate the similarity of distinct populations and to classify units of unknown origin to known populations. The discriminant technique is also used in genetic studies in order to summarize the genetic differentiation between groups. In this kind of studies with Single Nucleotide Polymorphism (SNP) or re-sequencing data sets, usually the number of variables (alleles) is greater than the number of observations (units), so discriminant methods available for data sets with more variables than units are necessary. Furthermore, the class prediction is currently one of the most important tasks in biomedical studies. The diagnosis of diseases, as cancer type or psychiatric disorder, has recently received a great deal of attention. With actual data, classification presents serious difficulties, because diagnosis is based on both clinical/pathological features (usually nominal data) and gene expression information (continuous data). For this reason, classification rules that could be applied to all types of data are desirable. The most popular classification rules are the linear (LDA) and quadratic (QDA) discriminant analyses (Fisher, 1936), which are easy to use as they are found in most statistical packages. However, they require the assumption of normally distributed data; when this condition is violated, they may lead poor classification results. Many distinct classifiers exist, differing in the definition of the classification rule and they utilize statistical (Golub et al., 1999; Hastie et al., 2001) or machine learning (Breiman, 2001; Boulesteix et al., 2008) methods. However, the problem of classification with data obtained from microarrays is challenging because there are a large number of genes and a relatively small number of samples. In this situation, the classification methods based on the within-class covariance matrix failed as its inverse is not defined. This is known as the singularity or under sample problem (Krzanowski et al., 1995). The shrunken centroid method that can be seen as a modification of the diagonal discriminant analysis (Dudoit et al., 2002) was developed for continuous high-dimensional data (Tibshirani et al., 2002). Nowadays, another issue that requires attention is the class-unbalanced situation, that is, the number of units belonging to each class is not the same. Some classifiers on class-unbalanced data tend to classify most of the new data in the majority class. This bias is higher when using high dimensional data. Recently, a method which improves the shrunken centroid method when the high-dimensional data is class-unbalanced was presented (Blagus and Lusa,

2013). Furthermore, some statistical approaches are characterized by having an explicit underlying probability model, but it is not possible to assume always this requirement. One of the most popular nonparametric machine learning classification method is the k-nearest neighbor classification (k-NN) (Cover and Hart, 1967; Duda et al., 2000). Given a new unit to be classified, this method finds the k nearest neighbors and classifies the new unit in the class to which belong the majority of neighbours, but the classification may depend on the k selected value. As ecologists have repeatedly argued, the Euclidean distance is inappropriate for raw species abundance data involving null abundances (Orloci, 1967; Legendre and Legendre, 1998) and it is necessary to use discriminant analyses that incorporate adequate distances. In this situation, discriminant analysis based on distances (DB-discriminant), where any symmetric distance or dissimilarity function can be used, is a useful alternative (Cuadras, 1989, 1992; Cuadras et al., 1997; Anderson and Robinson, 2003). To our knowledge, this technique is only included in GINGKO a suite of programs for multivariate analysis, oriented towards ordination and classification of ecological data (De Caceres et al., 2003; Bouin, 2005; Kent, 2011). These programs are written in Java language, so it is therefore necessary to have a Java Virtual Machine to execute it. Even though GINGKO is a very useful tool, it does not provide the option of a class prediction for new unlabeled units or feature's selection. Recently, data depth was proposed as the basis for nonparametric classifiers (Jornstein, 2004; Ghosh and Chaudhuri, 2005; Jin and Cui, 2010; Hlubinka and Vencalek, 2013). A depth of a unit is a nonnegative number, which measures the centrality of the unit. That is, depth in the sample version reflects position of the unit with respect to the observed data cloud. The so-called maximal depth classifier is the simple and natural classifier defined from a depth function: to allocate a new observation to the class to which it has maximal depth. There are many possibilities how to define the depth of the data (Liu, 1990; Vardi and Zhang, 2000; Zuo and Serfling, 2000; Serfling, 2002), nevertheless the computation of the most popular depth functions is very slow, in particular for high dimensional data and the time needed for classification grows rapidly. A new less-computer intensive depth function I (Irigoien et al., 2013a) was developed, but the authors did not study their use in relation to the classification problem.

A discriminant method should have several abilities. First, the classifier rule has to be able to properly separate the classes. In this sense, the classifier evaluation is most often based on the error rate, the percentage of incorrect prediction divided by the total number of predictions. Second, the rule has to be useful to classify new unlabeled units. Then, cross validation evaluation is needed. Cross-validation involves a series of sub-experiments, each of which involves the removal of a subset of objects from a data set (the test set), construction of a classifier using the remaining objects in the data set (the model building set), and subsequent application of the resulting model to the removed objects. Leave-one-out method is a special case of cross-validation, and it considers each single object in the data set as a test set. Furthermore, other measures as the sensitivity, specificity, positive predictive value for each class and the generalized correlation coefficient are useful to known the ability of the rule in the prediction task.

Here we introduce **WeDiBaDis**, an R package which provides a user-friendly interface to run the DB-discriminant analysis and a new classification procedure, the weighted-distance-based discriminant (WDB-discriminant) that performs well and improves the DB-discriminant rule. It is based on both, the DB-discriminant rule and the depth function I (Irigoien et al., 2013a). First, we will describe the DB and WDB discriminant rules. Then, we will provide details about **WeDiBaDis** package and we will illustrate its use and its main outputs using an ecological and a genetic data sets. To compare both DB and WDB rules and in order to avoid the criticism that artificial data can favour particular methods, we present a large analysis of thirty-eight high-dimensional class-unbalanced cancer gene expression data sets, three of them including clinical features. Furthermore, the data sets include more than two classes. Finally, we conclude the paper presenting the main conclusions. **WeDiBaDis** is available at `https://github.com/ItziarI/WeDiBaDis`.

Discriminant rules and evaluation criteria

Let $\mathbf{y}_i$ ($i = 1, 2, ..., n$) be m-dimensional units measured in any kind of features, with associated class labels $L_i \in \{1, 2, \ldots, K\}$, where n and K denote the number of units and classes, respectively. Let $\mathbf{Y}$ be the matrix of all units and d a distance defined between any pair of units, $d_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$. Let $\mathbf{y}^*$ a new unlabeled unit to be classified in one of the given classes C_k, $k = 1, 2, ..., K$.

DB-discriminant

The distance-based or DB-discriminant rule (Cuadras et al., 1997) takes as discriminant score

$$\delta_k^1(\mathbf{y}^*) = \hat{\phi}^2(\mathbf{y}^*, C_k), \tag{1}$$

where $\hat{\phi}^2(\mathbf{y}^*, C_k)$ is the proximity function which measures the proximity between $\mathbf{y}^*$ and C_k. This function is defined by,

$$\hat{\phi}^2(\mathbf{y}^*, C_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} d^2(\mathbf{y}^*, \mathbf{y}_i) - \frac{1}{2n_k^2}\sum_{i,j=1}^{n_k} d^2(\mathbf{y}_i, \mathbf{y}_j), \tag{2}$$

where n_k indicates the number of units in class k. Note that the second term in (2),

$$\hat{V}(C_k) = \frac{1}{2n_k^2}\sum_{i,j=1}^{n_k} d^2(\mathbf{y}_i, \mathbf{y}_j),$$

called geometric variability of C_k, measures the dispersion of C_k. When d is the Euclidean distance, $\hat{V}(C_k)$ is the trace of the covariance matrix of $\mathbf{Y}$.

The DB classification rule allocates $\mathbf{y}^*$ to the class which it has minimal proximity value:

$$C_{DB}(\mathbf{y}^*) = L \quad \text{where} \quad \delta_L^1(\mathbf{y}^*) = \min_{k=1,\ldots,K}\left\{\delta_k^1(\mathbf{y}^*)\right\}. \tag{3}$$

That is, this distance-based rule assigns a unit to the nearest group. Furthermore, using appropriate distances (3) reduces to some classic and well studied rules (see Table 1 in Cuadras et al. (1997)). For example, under the normality assumption, (3) is equivalent to linear discriminant or to quadratic discriminant if the Mahalanobis distance or the Mahalanobis distance plus a constant is selected, respectively.

WDB-discriminant

For any unit $\mathbf{y}$, let I_k be the depth function in class C_k defined by (Irigoien et al., 2013a),

$$I_k(\mathbf{y}) = \left[1 + \frac{\hat{\phi}^2(\mathbf{y}, C_k)}{\hat{V}(C_k)}\right]^{-1}. \tag{4}$$

Function I takes values in $[0, 1]$ and it verifies the following desirable properties: For a distribution having a uniquely defined "center" I attains maximum value at this center (maximality at center); When one unit moves away from the deepest unit (the unit at which the depth function attains maximum value; in particular, for a symmetric distribution, the center) along any fixed ray through the center, the depth at this unit decreases monotonically (monotonicity relative to the deepest point) and the depth of a unit $\mathbf{y}$ should approach zero as $||\mathbf{y}||$ approaches infinity (vanishing at infinity). According to the distance used, the depth of a unit may depend on or not of the underlying coordinate system or, in particular, of the scales of the underlying measurements. In any case the affine invariance is hold for translations and rotations. Thus, according to Zuo and Serfling (2000), I is a *type C* depth function. As I is a depth function, it assigns to any observation a degree of centrality. While most of the depth functions assign zero depth to units outside a convex hull and then, it is possible that some training units have zero depth, function (4) attains the zero value if $V(C_k) = 0$, that is, in presence of a constant distribution.

For each class C_k we weighted the discriminant score δ_k^1 by $1 - I_k(\mathbf{y}^*)$, that is, given a new unit $\mathbf{y}^*$, we define a new discriminant score for class k by:

$$\delta_k^2(\mathbf{y}^*) = \delta_k^1(1 - I_k(\mathbf{y}^*)) = \phi^2(\mathbf{y}^*, C_k)(1 - I_k(\mathbf{y}^*)). \tag{5}$$

The shrinkage we use, reduces the proximity values, being this reduction greater for deepest units. Thus, this new classification rule,

$$C_{WDB}(\mathbf{y}^*) = L \quad \text{where} \quad \delta_L^2(\mathbf{y}^*) = \min_{k=1,\ldots,K}\left\{\delta_k^2(\mathbf{y}^*)\right\}, \tag{6}$$

allocates a new unit $\mathbf{y}^*$ to the class which it has minimal proximity and maximal depth values.

Evaluation criteria

First consider the case of two classes ($K = 2$) and the most common measures of performance for a classification rule. As it is usual in medical statistics and for a fixed class k, let TP, FN, FP and TN denote the true positive (number of units of class k correctly classified in class k), the false negative (number of units of class k misclassified as units in class l, with $l \neq k$), the false positive (number of units of class l, with $l \neq k$ misclassified as units in class k) and the true negative (number of units of

class l, with $l \neq k$ correctly classified as units in class l), respectively. Then (Zhou et al., 2002), the sensitivity (recall) for class k is defined as the ability of a rule to classified correctly units belonging to class k, thus $Q_k^{se} = \frac{TP}{TP+FN}$. The specificity is the ability of a rule to correctly exclude a unit from class k when it really belongs to another class, thus $Q_k^{sp} = \frac{TN}{TN+FP}$. Furthermore, the positive predictive value (precision) is the probability that a classification in class k is correct, thus $P_k^+ = \frac{TP}{TP+FP}$ and the negative predictive value is the probability that a classification in class l with $l \neq k$ is correct, thus $P_k^- = \frac{TN}{TN+FN}$. However, these measures do not take into account all the TP, FN, FP and TN values. For this reason, in biomedical applications it is often used the Matthew's correlation coefficient (Matthews, 1975), MC, which is defined by:

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

It ranges from -1 if all the classifications are wrong to +1 for perfect classification. Null values indicate that the classifications are random or the classifier always predicts only one of the two classes.

In the general case of K classes with $K \geq 2$, one obtains a $K \times K$ contingency or confusion matrix $\mathbf{Z} = (z_{kl})$, where z_{kl} is the number of times that units are classified to be in class l while belonging in reality to class k. Then, $z_{k.} = \sum_l z_{kl}$ and $z_{.l} = \sum_k z_{kl}$ represent the number of units belonging to class k and the number of units predicted to be in class l, respectively. Obviously $n = \sum_{kl} z_{kl} = \sum_k z_{k.} = \sum_l z_{.l}$. One standard criterium to evaluate a classification rule is to compute the percentage of all correct predictions,

$$Q_t = 100\frac{\sum z_{kk}}{n}, \tag{7}$$

the percentage of units correctly predicted to belong to class k relative to the total number of units in class k (sensitivity for class k),

$$Q_k^{se} = 100\frac{z_{kk}}{z_{k.}}, \tag{8}$$

the percentage of units correctly predicted to belong to any class l with $l \neq k$ relative to the total number of units in any class l with $l \neq k$ (specificity of class k),

$$Q_k^{sp} = 100\frac{\sum_{l\neq k} z_{l.} - \sum_{l\neq k} z_{lk}}{n - z_{k.}}, \tag{9}$$

and the percentage of units correctly classified to be in class k with respect to the total number of units classified in class k (positive predictive value for class k),

$$P_k^+ = 100\frac{z_{kk}}{z_{.k}}. \tag{10}$$

However, we also considered a generalization of the Matthew's correlation coefficient, the so called generalized squared correlation GC^2 (Baldi et al., 2000), which is defined by

$$GC^2 = \frac{\sum_{k,l} (z_{kl} - e_{kl})^2/e_{kl}}{n(K-1)}, \tag{11}$$

where $e_{kl} = \frac{z_{k.}z_{.l}}{n}$. This coefficient ranges between 0 and 1, and may often provide a much more balanced evaluation of the prediction than, for instance, the above percentages. A value equal to zero indicates that there is at least one class in which no units are classified.

Another interesting coefficient is the Kappa statistics which measures the agreement of classifiaction to the true class (Cohen, 1960; Landis and Koch, 1977). It can be calculated by:

$$Kappa = \frac{\frac{TP+TN}{n} - \frac{(TN+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{n^2}}{1 - \frac{(TN+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{n^2}},$$

and the interpretation is: $Kappa < 0$ less than chance agreement; $Kappa$ in $0.01 - 0.20$ slight agreement; $Kappa$ in $0.21 - 0.40$ fair agreement; $Kappa$ in $0.41 - 0.60$ moderate agreement; $Kappa$ in $0.61 - 0.80$ substantial agreement; $Kappa$ in $0.81 - 0.99$ almost perfect agreement.

Finally, another measure used as a result of classification is the F_1 statistic (Powers, 2011). For each class, it is calculated based on the precision P_k^+ and the recall Q_k^{se} as follows: $F_1 = 2 * \frac{P_k^+ Q_k^{se}}{P_k^+ + Q_k^{se}}$. However, note that F_1 does not take the true negatives into account.

Distance functions

The DB and WDB procedures require the previous calculation of a distance between units. In biomedical, genetic and ecological studies different types of dissimilarities are frequently used, for this reason, **WeDiBaDis** includes some distance functions. Although these distances can be found in other packages they were included for ease of use to non-expert R users. Thus the package contains the usual Euclidean distance,

$$d_E(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{\sum_{k=1}^{m} (y_{ik} - y_{jk})^2}, \tag{12}$$

the well known correlation distance, where r is the Pearson correlation coefficient,

$$d_c(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{(1 - r(\mathbf{y}_i, \mathbf{y}_j))}, \tag{13}$$

and the Mahalanobis distance (Mahalanobis, 1936) with S the variance-covariance matrix,

$$d_M(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{(\mathbf{y}_i - \mathbf{y}_j)' S^{-1} (\mathbf{y}_i - \mathbf{y}_j)}. \tag{14}$$

The function named `mahalanobis()` that calculates the Mahalanobis distance already exists in the `stats` package, but it is not suitable in our context. While this function calculates the Mahalanobis distance with respect to a given center, our function is designed to calculate the Mahalanobis distance between each pair of units given a data matrix.

Next, we briefly comment the other distances included in the package. The Bhattacharyya distance (Bhattacharyya, 1946) is a very well-known distance between populations in the genetic context. Each population is characterized by a vector $(p_{i1}, ..., p_{im})$ whose coordinates are the relative frequencies of the features (usually chromosomal arrangements), with

$$p_{ij} > 0, j = 1, ..., m \quad \text{and} \quad \sum_{j=1}^{m} p_{ij} = 1, i = 1, ..., n.$$

Then, the distance between two units (populations) with frequencies $\mathbf{y}_i = (p_{i1}, ..., p_{im})$ and $\mathbf{y}_j = (p_{j1}, ..., p_{jm})$ is defined by:

$$d_B(\mathbf{y}_i, \mathbf{y}_j) = \arccos \sum_{l=1}^{m} \sqrt{p_{il} p_{jl}}. \tag{15}$$

The Gower distance (Gower, 1971), used for mixed variables, is defined by:

$$d_G(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{2(1 - s(\mathbf{y}_i, \mathbf{y}_j))}. \tag{16}$$

with $s(\mathbf{y}_i, \mathbf{y}_j)$ is the similarity coefficient between unit $\mathbf{y}_i = (\mathbf{x}_i, \mathbf{q}_i, \mathbf{b}_i)$ and unit $\mathbf{y}_j = (\mathbf{x}_j, \mathbf{q}_j, \mathbf{b}_j)$, where $\mathbf{x}_\cdot, \mathbf{q}_\cdot, \mathbf{b}_\cdot$ are the values for the m_1 continuous, the m_2 binary and the m_3 qualitative features, respectively. Coefficient $s(\mathbf{y}_i, \mathbf{y}_j)$ is calculated as:

$$s(\mathbf{y}_i, \mathbf{y}_j) = \frac{\sum_{l=1}^{m_1} \left(1 - \frac{|\mathbf{x}_{il} - \mathbf{x}_{jl}|}{R_l}\right) + a + \alpha}{m_1 + (m_2 - d) + m_3},$$

with R_l the range of the lth continuous variable ($l = 1, ..., m_1$); for the m_2 binary variables, a and d represent the number of matches presence-presence and absence-absence, respectively and α is the number of matches between states for the m_3 qualitative variables. Note that there is also the `daisy()` function in `cluster` package, which can calculate the Gower distance for mixed variables. The difference between this function and `dGower()` in **WeDiBaDis** is that in `daisy()` the distance is calculated as $d(\mathbf{y}_i, \mathbf{y}_j) = 1 - s(\mathbf{y}_i, \mathbf{y}_j)$ and in `dGower()` as $d(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{2(1 - s(\mathbf{y}_i, \mathbf{y}_j))}$. Moreover, `dGower()` allows us to include missing values (such as NA) and therefore calculates distances based on Gower's weighted similarity coefficients. The `dGower()` function improves the function `dgower()` included in package **ICGE** (Irigoien et al., 2013b).

The Bray-Curtis distance (Bray and Curtis, 1957) is one of the most well-known ways of quantifying the difference between samples when the information is ecological abundance data collected at

different sampling locations, and it is defined by:

$$d_B(\mathbf{y}_i, \mathbf{y}_j) = \frac{\sum_{l=1}^{m} |y_{il} - y_{jl}|}{y_{i+} + y_{j+}}, \tag{17}$$

where y_{il}, y_{jl} are the abundance of specie l in samples i and j, respectively, and y_{i+}, y_{j+} are the total specie's abundance in samples i and j, respectively. This distance can be also found in the **vegan** package.

The Hellinger (Rao, 1995) and Orloci or chord distance (Orloci, 1967) distances are also measures recommended for quantifying differences between sampling locations when the ecological abundance of species is collected. The Hellinger distance is given by:

$$d_H(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{\sum_{l=1}^{m} \left(\sqrt{\frac{y_{il}}{\sum_{k=1}^{m} y_{ik}}} - \sqrt{\frac{y_{jl}}{\sum_{k=1}^{m} y_{jk}}} \right)^2}, \tag{18}$$

and the Orloci distance that represents the Euclidean distance computed after scaling the site vectors to length 1 is defined by:

$$d_O(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{\sum_{l=1}^{m} \left(\frac{y_{il}}{\sqrt{\sum_{k=1}^{m} y_{ik}^2}} - \frac{y_{jl}}{\sqrt{\sum_{k=1}^{m} y_{jk}^2}} \right)^2}. \tag{19}$$

This distance between two sites is equivalent to the length of a chord joining two points within a segment of a hypersphere of radius 1.

The Prevosti distance (Prevosti et al., 1975) is a very useful genetic distance between units representing populations. Now, we consider that genetic data is stored in a table where the rows represent the populations and the columns represent potential allelic states grouped by loci. The distance between two units at a single locus k with $m(k)$ allelic states is:

$$d_P(\mathbf{y}_i, \mathbf{y}_j) = \frac{1}{2\nu} \sum_{k=1}^{\nu} \sum_{s=1}^{m(k)} |p_{iks} - p_{jks}|, \tag{20}$$

where ν is the number of loci or chromosomes (in the case of chromosomal polymorphism) considered and p_{iks}, p_{jks} are the sample relative frequencies of the allele or chromosomal arrangement s in the locus or chromosome k, in the ith and jth population, respectively. With presence / absence data coded by 1 and 0, respectively, the term $\frac{1}{2\nu}$ will be omitted.

As we explain in the next section, **WeDiBaDis** allows to introduce any other distance selected by the user by menas of the distance matrix. Therefore, the user can work with any distance matrix that considers appropriate for their data set and analysis. For this reason, no more distances were included in our package.

Using the package

We have developed the WeDiBaDis package to implement both, the DB-discriminant and the new WDB-discriminant. It can be used with different distances and NA values are allowed. When an unit has a NA value in some features, those features are excluded in the computation of the distances for that unit and the computation is scaled up to the number m of features involved in the data set. Package **WeDiBaDis** requires the R version 3.3.1 or a greater version.

The principal function is `WDBdisc`,

```
WDBdisc(data,datatype,classcol,new.ind,distance,type,method)
```

For its arguments, one should specify:

- `data` a data matrix or a distance matrix. If the Prevosti distance will be used, `data` must be a named matrix where the name of the loci and allele must be separeted by a dot (Loci-Name.AlleleName)
- `datatype` if the data is a data matrix, `datatype="m"`; if the data is a distance matrix `datatype="d"`

- `classcol` a number indicating which column in the data contains the class variable. By default the class variable is in the first column
- `new.ind` only required if there are new unlabeled units to be classified; if `datatype="m"` it is a matrix containing the feature values for the new units to be classified; if `datatype="d"` it is a matrix containing the distances between the new units to be classified and the units in the classes
- `distance` the distance measure to be used. This must be one of "euclidean" (default option), "correlation" , "Bhattacharyya", "Gower", "Mahalanobis", "BrayCurtis", "Orloci", "Hellinger", and "Prevosti".
- `type` only required if `distance = "Gower"`. `type` is a list object `type = list(cuant,nom,bin)` indicating the position of the columns for continuous (`cuant`), nominal (`nom`) and binary (`bin`) features, respectively
- `method` the discriminant method to be used. This must be one of "DB", "WDB" for the DB-discriminant and WDB-discriminant, respectively. The method by default is WDB

The function returns an object with associated `plot` and `summary` methods offering:

- The classification table obtained with the leave-one-out cross-validation
- The total well classification rate in percentage (Q_t)
- The generalized squared correlation (GC^2)
- The sensitivity, specificity and positive predictive values for each class (Q_k^{se}, Q_k^{sp} and P_k^+, respectively)
- The Kappa and F_1 statistics
- The assigned class for new unlabeled units to be classified
- A barplot for the classification table
- A barplot for the sensitivity, specificity and positive predictive values for each class

Moreover, given a data set, the commented distances in Distance funtions Section can be obtained through the functions: `dcor` (correlation distance); `dMahal` (Mahalanobis distance); `dBhatta` (Bhattacharyya distance); `dGower` (Gower distance); `dBrayCurtis` (Bray and Curtis distance); `dHellinger` (Hellinger distance); `dOrloci` (Orloci distance) and `dPrevosti` (Prevosti distance).

Example 1: Ecological data

We consider the data from Fielding (2007), which relate to the core area (the region close to the nest) of the golden eagle *Aquila chrysaetos* in three regions of Western Scotland. The data consist of eight habitat variables: POST (mature planted conifer forest in which the tree canopy has closed); PRE (pre-canopy closure planted conifer forest); BOG (flat waterlogged land); CALL (Calluna (heather) heath land); WET (wet heath, mainly purple moor grass); STEEP (steeply sloping land); LT200 (land below 200 m), and L4-600 (land between 200 and 400 m). The values are the numbers of four-hectare grid cells covered by the habitat, whose values are the amounts of each habitat variable, measured as the number of four hectare blocks within a region defined as a "core area". In order to evaluate if the habitat variables allow to discriminate between these three regions we performed for example, a WDB-discriminant using the Euclidean distance using the following instructions,

```
library(WeDiBaDis)
out<-WDBdisc(data=datafile,datatype="m",classcol=1)
summary(out)
```

The `summary` method shows, as usual, the following more complete information,

```
Discriminant method: WDB
Leave-one-out confusion matrix:
           Predicted
 Real    1     2     3
 1       7     0     0
 2       0    14     2
 3       2     0    15
Total correct prediction:   90%
Generalized squared correlation:  0.7361
Cohen's Kappa coefficient:  0.84375
```

```
Sensitivity for each class:
 1       2      3
100.00 87.50  88.24
Predictive value for each class:
 1      2       3
77.78 100.00  88.24
Specificity for each class:
  1     2     3
87.88 91.67 91.30
F1-score for each class:
    1     2     3
87.50 93.33 88.24
------  ------  ------  ------  ------  ------
No predicted individuals
```

As we can observe, perfect classification is obtained for samples from region 1. For regions 2 and 3, only two samples were not correctly classified.

If we want to obtain the barplot for the classification table (see Figure 1), we will use the sentence

```
plot(out)
```

The next commands:

```
outplot <- summary(out, show=FALSE)
plot(outplot)
```

generate the sensitivity, specificity and positive predicted values barplot (see Figure 2).

Finally, if the user wants to perform a DB discriminant using a different distance that the Euclidean, the following sentences are needed,

```
library(WeDiBaDis)
out<-WDBdisc(data=datafile,datatype="m",distance="name of the distance",method="DB",classcol=1)
summary(out)
plot(out)
outplot <- summary(out, show=FALSE)
plot(outplot)
```

Example 2: population genetics data

The chromosomal polymorphism for inversions is very useful to characterize the natural populations of *Drosophila subobscura*. Furthermore, lethal genes located in chromosomal inversions allowed to understand important evolutionary events. We consider the data from a study of 40 samples of this polymorphism for the O chromosome of this species (Solé et al., 2000; Balanyà et al., 2004; Mestres et al., 2009). Four groups can be considered: NLE with 16 no lethal European samples, LE with 4 lethal European samples, NLA with 14 no lethal American samples and LA with 6 lethal American samples. In this example, two samples one of the group NLA and one of the group NLE were randomly selected, and considered as new unlabeled units to be classified. The Bhattacharyya distances between all pairs of units were calculated. Therefore, the input for function `WDBdisc` is an $n \times (n+1)$ matrix `dat`=$(l_i,\ d_B(\mathbf{y}_i,\mathbf{y}_1),\ldots,d_B(\mathbf{y}_i,\mathbf{y}_n))_{i=1,\ldots,n}$ where the first column contains the class label and the followings columns the distance matrix. Furthermore, `xnew` is a two row matrix where each row contains the distances between the new unlabeled units to be classified and the units in the four classes. In this situation, if the WDB procedure is selected in order to classify the `xnew` units, and to obtain the available graphics in the package, we will use,

```
library(WeDiBaDis)
out<-WDBdisc(data=dat,datatype="d",classcol=1,new.ind=xnew)
summary(out)
plot(out)
outplot <- summary(out, show=FALSE)
plot(outplot)
```

The `summary` shows the following information and we can see that the `xnew` units were correctly classified:

```
Discriminant method: WDB
Leave-one-out confusion matrix:
        Predicted
 Real LA  LE  NLA  NLE
  LA   6   0   0  0
  LE   0   3   0  1
  NLA  0   0  13  0
  NLE  0   3   0 12
Total correct prediction:   89.47%
Generalized squared correlation:  0.7442
Cohen's Kappa coefficient:  0.8509804
Sensitivity for each class:
 LA      LE     NLA      NLE
100.00  75.00  100.00   80.00
Predictive value for each class:
 LA      LE     NLA      NLE
100.00  50.00  100.00   92.31
Specificity for each class:
LA      LE     NLA      NLE
87.50 91.18   84.00    95.65
F1-score for each class:
    LA     LE    NLA    NLE
100.00  60.00 100.00  85.71
------ ------ ------ ------ ------ ------
Prediction for new individuals:
Pred. class
1 "NLE"
2 "NLA"
```

Now, the two unlabeled new units were correctly classified and the barplots are in Figure 3 and Figure 4, respectively.

Data files

The package contains some examples of data files with the corresponding explanation. The data sets are `corearea` containing the data for the example presented in the subsection Example 1: Ecological data; `abundances` which is a simulated data set for abundance data matrix, and `microsatt` a data set containing allele frequencies for 18 cattle breeds (bull or zebu), of French and African descent, typed on 9 microsatellites.

Computing time

To illustrate the time consumed by the WDB procedure, which requires more computation than DB, we performed the following simulation with artificial data. We generated multinormal samples containing 50, 100, 200, 300,...,900, 1000, 2000 and 3000 units, respectively. Then, for each sample size we created sets containing respectively 50, 100, 500, 1000, 1500, 2000, 2500, ..., 4500 and 5000 features. For each of the above combination of sample size and features, we considered 2, 3, 4 and 10 classes. All the computations presented in this paper have been performed on a Personal Computer Intel(R) Core(TM) i5-2450M with 6 GB of memory using a single 2.50GHz CPU processor. The results of the simulation for two classes are displayed in Figure 5, where the elapsed time (the actual elapsed time since the process started) are reported in seconds. We can observe that the runtime is mainly affected by the number of units (Figure 4, top), but very little by the number of variables (Figure 4, bottom). This is expected since the procedure is based on distances and therefore the dimension of the distance matrix (number of units) determines the runtime required. The number of classes also affects the runtime, although its variation with increasing the number of classes is very slight. For example, with 300 units and 4000 variables, the elapsed time for 2, 3, 4 and 10 classes are 3.38, 3.40, 3.62 and 4.82 seconds, respectively.

DB and WDB comparison using cancer data sets

In order to compare the performance of DB and WDB procedures, thirty-eight available cancer data sets were considered in our analysis (Table 1), which are available at http://bioinformatics.rutgers.

edu/Static/Supplements/CompCancer/datasets.htm and Lê Cao et al. (2010). As we can observe (Table 1) three of them include clinical features and some of the data sets presents unbalanced classes. We performed the evaluation for DB and WDB classifiers using the leave-one-out procedure. We present the total misclassification rate $MQ_t = 100 - Q_t$ and the generalized squared correlation coefficient GC^2 (Table 2). For simplicity, the sensitivity $Q_k i^{se}$, the specificity Q_k^{sp}, the positive predictive value P_k^+ for each class, the Kappa and F_1 statistcis are no presented. For the microarray data sets with only continuous features we used the Euclidean distance, and for those including clinical and genetic data, we considered the Gower distance (Gower, 1971). As we can observe in Table 2, considering only MQ_t, the total misclassification percentage rate, WDB was the best classifier in 18 data sets and it shared this quality in 11 data sets with DB (Wilcoxon signed rank test; one side p-value = 0.0265). Using the generalized squared correlation GC^2 coefficient (Table 2), WDB was the best rule in 16 data sets and it shared this quality in 11 data sets with DB (Wilcoxon signed rank test; one side p-value = 0.0378). Note that for data sets 30 and 38 the GC^2 value is 0. For example, in the Risinger-2003 case, all units of the second class (class with 3 units) were badly classified with DB and WDB methods. However, while with the DB method, 4 units belonging to other classes were badly classified in class 2, with the WDB method none of the units of other classes were badly classified in class 2, and for this reason the GC^2 is equal to 0. With the Yeoh-2002-v2 data set something similar happened. For all these results, WDB seems to obtain in general the best results and to be a slightly better in the case where classes are unbalanced with respect to their sizes.

Conclusions

The package **WeDiBaDis**, available at this momnet in https://github.com/ItziarI/WeDiBaDis, is the implementation of two discriminant analysis procedures in an R environment. The classifiers are the Distance-Based (DB) and the new proposed procedure Weighted-Distance-Based (WDB), which are useful to solve the classification problem for high-dimensional data sets with mixed features, or when the input information is a distance matrix. This software provides functions to compute both discriminant procedures and to assess the performance of the classification rules it offers: the leave-one-out classification table; the general correlation coefficient; the sensitivity, specificity and positive predictive value for each class; the Kappa and the F_1 statistics.The package also presents these results in a graphical form (barplots for the classification table and, for sensitivity, specificity and positive predictive values, respectively). Furthermore, it allows the classification for new unlabeled units. **WeDiBaDis** provides a user-friendly environment, which can be of great utility in biology, ecology, biomedical, and in general, any applied study involving discrimination between groups and classification of new unlabeled units. In addition, it can be very useful in multivariate methods courses aimed at biologists, medical researches, psychologists etc.

Acknowledgements

This research was partially supported: II by the Spanish 'Ministerio de Economia y Competitividad' (TIN2015-64395-R) and by the Basque Government Research Team Grant (IT313-10) SAIOTEK Project SA-2013/00397 and by the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab). FM by the Spanish 'Ministerio de Economia y Competitividad' (CTM2013-48163) and by Grant 2014 SGR 336 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya. CA by the Spanish 'Ministerio de Economia y Competitividad' (SAF2015-68341-R), by the Spanish 'Ministerio de Economia y Competitividad' (TIN2015-64395-R) and by Grant 2014 SGR 464 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya.

Bibliography

M. J. Anderson and J. Robinson. Generalized discriminant analysis based on distances. *Australian and New Zealand Journal of Statistics*, 45:301–318, 2003. [p2]

I. Balanyà, E. Solé, J. M. Oller, D. Sperlich, and L. Serra. Long-term changes in chromosomal inversion polymorphism of *d. subobscura*: Ii. european populations. *Journal of Zoological Systematics and Evolutionary Research*, 42:191–201, 2004. [p8]

P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000. [p4]

A. Bhattacharyya. On a measure of divergence of two multinominal populations. *Sankhya*, 7:401–406, 1946. [p5]

R. Blagus and L. Lusa. Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:64, 2013. [p1]

G. Bouin. Computer program review: Ginkgo, a multivariate analysis package. *Journal of Vegetation Science*, 16:355–359, 2005. [p2]

A. L. Boulesteix, C. Porzelius, and M. Daumer. Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value. *Bioinformatics*, 24:1698–1706, 2008. [p1]

J. R. Bray and J. T. Curtis. An ordination of upland forest communities of southern wisconsin. *Ecological Monographs*, 27:325–349, 1957. [p5]

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. [p1]

J. Cohen. A coefficient of agreement for nominal scales. *Educational anp Psychological Measurement*, 20: 37–46, 1960. [p4]

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967. [p2]

C. M. Cuadras. *Statistical Data Analysis and Inference*, chapter Distance Analysis. In: Discrimination and Classfication Using both Continuous and Categorical Variables, pages 459–473. Elsevier Science Publishers BV, Amsterdam, 1989. [p2]

C. M. Cuadras. Some examples of distance based discrimination. *Biometrical Letters*, 29:3–20, 1992. [p2]

C. M. Cuadras, J. Fortiana, and F. Oliva. The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification*, 14:117–136, 1997. [p2, 3]

M. De Caceres, F. Oliva, and X. Font. Ginkgo, a multivariate analysis program oriented towards distance-based classifications. In *International Conference on Correspondence Analysis and Related Methods (CARME' 03)*, 2003. [p2]

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interescience Publication. John Wiley and Sons, New York, 2000. [p2]

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association*, 97:77–87, 2002. [p1]

A. H. Fielding. *Cluster and Classification Techniques for the Biosciences*. University Press, Cambridge, 2007. [p7]

R. A. Fisher. The use of multiple measurements in taxonomic problems. *The Annals of Eugenics*, 7: 179–188, 1936. [p1]

A. K. Ghosh and P. Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11:1–27, 2005. [p2]

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. [p1]

J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–871, 1971. [p5, 10]

T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome Biology*, 2:1–12, 2001. [p1]

D. Hlubinka and O. Vencalek. Depth-based classification for distributions with nonconvex support. *Journal of Probability and Statistics*, 28:1–7, 2013. [p2]

I. Irigoien, F. Mestres, and C. Arenas. The depth problem: Identifying the most representative units in a data group. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:161–172, 2013a. [p2, 3]

I. Irigoien, B. Sierra, and C. Arenas. Icge: An r package for detecting relevant clusters and atypical units in gene expression. *BMC Bioinformatics*, 13:30–41, 2013b. [p5]

J. Jin and H. Cui. Discriminant analysis based on statistical depth. *Journal of Systems Science and Complexity*, 23:362–371, 2010. [p2]

R. Jornstein. Clustering and classification based on l_1 data depth. *Journal of Multivariate Analysis*, 90: 67–89, 2004. [p2]

M. Kent. *Vegetation Description and Data Analysis: A Practical Approach*. Wiley-Blackwey, 2011. [p2]

W. I. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas. Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995. [p1]

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977. [p4]

K. A. Lê Cao, E. Meugnier, and G. J. McLachlan. Integrative mixture of experts to combine clinical factors and gene markers. *Bioinformatics*, 26:1192–1198, 2010. [p10]

P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, Amsterdam, 1998. [p2]

R. Y. Liu. On a notion of data depth based on random simplices. *Annals of Statistics*, 18:405–414, 1990. [p2]

P. V. Mahalanobis. On the generalized distance in statistics. *Procedures of the Natural Institute of Science of India*, 2:49–55, 1936. [p5]

B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica Biophysica Acta*, 405:442–451, 1975. [p4]

F. Mestres, J. Balanyà, M. Pascual, C. Arenas, G. W. Gilchrist, R. B. Huey, and L. Serra. Evolution of chilean colonizing populations of *d. subobscura*: lethal genes and chromosomal arrangements. *Genetica*, 136:37–48, 2009. [p8]

L. Orloci. An agglomerative method for classification of plant communities. *Journal of Ecology*, 55: 193–205, 1967. [p2, 6]

D. M. W. Powers. Evaluation:from precision, recall and f-measure to roc, informedness, markedness and correlaction. *Journal of Machine Learning Technologies*, 2:37–63, 2011. [p4]

A. Prevosti, J. Ocaña, and G. Alonso. Distances between populations of *d. subobscura*, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics*, 45:231–241, 1975. [p6]

C. R. Rao. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Qüestiió*, 19:23–63, 1995. [p6]

R. Serfling. *Statistic and Data Analysis Based on L_1-Norm and Related Methods*, chapter A Depth Function and a Scale Curve Based on Spatial Quantiles, pages 25–38. Birkhäuser, Boston, 2002. [p2]

E. Solé, F. Mestres, J. Balanyà, C. Arenas, and L. Serra. Colonization of america by *d. subobscura*: Spatial and temporal lethal-gene allelism. *Hereditas*, 133:65–72, 2000. [p8]

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99:6567–6572, 2002. [p1]

Y. Vardi and C. Zhang. The multivariate l_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97:1423–1426, 2000. [p2]

X. H. Zhou, N. A. Obuchowski, and D. K. McClish. *Statistical Methods in Diagnostic Medicine*. Wiley Series in Probability and Statistics. John Wiley and Sons, New Jersey, 2002. [p4]

S. Zuo and R. Serfling. General notions of statistical depth function. *Annals of Statistics*, 28:461–482, 2000. [p2, 3]

Itziar Irigoien
Department of Computation Science and Artificial Intelligence
University of the Basque Country
Donostia, Spain
itziar.irigoien@ehu.eus

Francesc Mestres
Department of Genetics, Microbiology and Statistics. Genetics Section
University of Barcelona
Barcelona, Spain
fmestres@ub.edu

Concepcion Arenas
Department of Genetics, Microbiology and Statistics. Statistics Section
University of Barcelona
Barcelona, Spain
carenas@ub.edu

ID	Data set	K	n	n_i	p	*cuant*	*quali*
1	Alizadeh-2000-v1	2	42	21(50%), 21(50%)	1095	1095	
2	Alizadeh-2000-v2	3	62	42(67.74%), 9(14.52%), 11(17.74%)	2093	2093	
3	Armstrong-2002-v1	2	72	24(33.33%), 48(66.67%)	1081	1081	
4	Armstrong-2002-v2	3	72	24(33.33%), 20(27.78%), 28(38.89%)	2194	2194	
5	Bhattacharjee-2001	5	203	139(68.47%), 17(8.37%), 6(2.96%), 21(10.34%), 20(9.85%)	1543	1543	
6	Bittner-2000-V1	2	38	19(50%), 19(50%)	2201	2201	
7	Bittner-2000-V2	3	38	19(50%), 12(31.58%), 7(18.42%)	2201	2201	
8	Breast	2	256	75(29.30%), 181(70.70%)	5545	5537	8
9	Bredel-2005	3	50	31(62%), 14(28%), 5(10%)	1739	1739	
10	Chen-2002	2	179	104(58.10%), 75(41.90%)	85	85	
11	Chowdary-2006	2	104	62(59.62%), 42(38.89%)	182	182	
12	CNS	2	60	21(35%), 39(65%)	7134	7128	6
13	Dyrskjot-2003	3	40	9(22.5%), 20(50%), 11(27.5%)	1203	1203	
14	Garber-2001	4	66	17(25.76%), 40(60.61%), 4(6.06%), 5(7.58%)	4553	4553	
15	Golub-1999-v1	2	72	47(65.28%), 25(34.72%)	1877	1877	
16	Golub-1999-v2	3	72	38(52.78%), 9(12.5%), 25(34.72%)	1877	1877	
17	Gordon-2002	2	181	31(17.13%), 150(82.87%)	1626	1626	
18	Khan-2001	4	83	29(34.94%), 11(13.25%), 18(21.69%), 25(30.12%)	1069	1069	
19	Laiho-2007	2	37	8(21.62%), 29(78.38%)	2202	2202	
20	Lapointe-2004-v1	3	69	11(15.94%), 39(56.52%), 19(27.54%)	1625	1625	
21	Lapointe-2004-v2	4	110	11(10%), 39(35.45%), 19(17.27%), 41(37.27%)	2496	2496	
22	Liang-2005	3	37	28(75.67%), 6(16.22%), 3(8.11%)	1411	1411	
23	Nutt-2003-v1	4	50	14(50%), 7(14%), 14(28%), 15(30%)	1377	1377	
24	Nutt-2003-v2	2	28	14(50%),14(50%)	1070	1070	
25	Nutt-2003-v3	2	22	7(31.82%),15(68.18%)	1152	1152	
26	Pomeroy-2002-v1	2	34	25(73.53%), 9(26.47%)	857	857	
27	Pomeroy-2002-v2	5	42	10(23.81%), 10(23.81%), 10(23.81%), 4(9.52%) 8(19.05%)	1379	1379	
28	Prostate	2	79	37(46.84%), 42(53.16%)	7892	7884	8
29	Ramaswamy-2001	14	190	11(5.79%), 11(5.79%), 20(10.53%), 11(5.79%), 30(15.79%), 11(5.79%), 22(11.28%), 11(5.79%), 10(5.26%),11(5.79%), 11(5.79%), 10(5.26%), 11(5.79%), 10(5.26%)	1363	1363	
30	Risinger-2003	4	42	13(30.95%), 3(7.14%), 19(45.24%), 7(16.67%)	1771	1771	
31	Shipp-2002-v1	2	77	58(75.32%), 19(24.67%)	798	798	
32	Singh-2002	2	102	50(49.02%), 52(50.98%)	339	339	
33	Su-2001	10	174	8(4.60%), 26(14.94%), 23(13.22%), 12(6.90%), 11(6.32%), 7(4.02%), 28(16.09%), 27(15.52%), 6(3.45%), 26(14.94%)	1571	1571	
34	Tomlins-2006-v1	5	104	27(25.96%), 20(19.23%), 32(30.77%), 13(12.5%), 12(11.54%)	2315 2315	2315 2315	
35	Tomlins-2006-v2	4	92	27(26.35%), 20(21.74%), 32(34.78%), 13(14.13%)	1288	1288	
36	West-2001	2	49	25(51.02%), 24 (48.98%)	1198	1198	
37	Yeoh-2002-v1	2	248	43(17.34%), 205(82.66%)	2526	2526	
38	Yeoh-2002-v2	6	248	15(6.05%), 27(10.89%), 64(25.81%), 20(8.06%), 43(17.34%), 79(31.85%)	2526	2526	

Table 1: Cancer data sets (ID = identification number). They present different number of classes (K), number of samples (n), number of samples in each class (n_i), number of features (p), number of continuous features (*cuant*) and number of qualitative features (*quali*). The percentage corresponding to the number of samples belonging to each class is in brackets in column five.

ID	$100 - Q_t$ DB	$100 - Q_t$ WDB	GC^2 DB	GC^2 WDB
1	**7.14**	**7.14**	**0.74**	**0.74**
2	1.61	**0.00**	0.94	**1.00**
3	8.33	**5.56**	0.684	**0.77**
4	**4.17**	**4.17**	**0.88**	**0.88**
5	19.21	**15.27**	0.49	**0.56**
6	**13.16**	**13.16**	**0.56**	**0.56**
7	**36.84**	**36.84**	**0.25**	**0.25**
8	32.81	**30.47**	0.11	**0.13**
9	**18.00**	**18.00**	**0.34**	**0.34**
10	11.17	**8.94**	0.61	**0.67**
11	18.27	**9.62**	0.42	**0.64**
12	41.67	**38.33**	**0.01**	**0.01**
13	15.00	**12.50**	0.58	**0.65**
14	**21.21**	28.79	**0.38**	0.19
15	6.94	**4.17**	0.72	**0.82**
16	**6.94**	**6.94**	**0.81**	**0.81**
17	**12.71**	13.26	**0.47**	0.42
18	**1.20**	**1.20**	**0.97**	**0.97**
19	**21.62**	**21.62**	**0.23**	**0.23**
20	31.88	**30.43**	0.23	**0.26**
21	**30.91**	**30.91**	**0.34**	**0.34**
22	13.51	**10.81**	0.72	**0.76**
23	**32.00**	34.00	**0.40**	0.33
24	17.86	**10.71**	0.43	**0.65**
25	**4.55**	9.09	**0.80**	0.67
26	29.41	**20.59**	0.12	**0.16**
27	**16.67**	21.43	**0.65**	0.63
28	**34.18**	**34.18**	**0.10**	**0.10**
29	36.84	**29.47**	0.44	**0.53**
30	28.57	**26.19**	**0.36**	0.00
31	29.87	**12.99**	0.24	**0.48**
32	**30.39**	**30.39**	**0.18**	0.16
33	20.11	**16.67**	0.63	**0.70**
34	**17.31**	21.15	**0.66**	0.58
35	**23.91**	26.09	**0.46**	0.41
36	20.41	**14.29**	0.35	**0.52**
37	**1.61**	2.02	**0.89**	0.87
38	**21.77**	24.60	**0.57**	0.00

Table 2: In the first column identification number for cancer data sets. In the second and third columns, total leave-one-out misclassification rate $100 - Q_t$ (in percentage) for classifiers DB and WDB, respectively. In bold the smallest misclassification rate. In the forth and fifth columns, generalized squared correlation GC^2 coefficient for classifiers DB and WDB, respectively. In bold the greater GC^2 value.

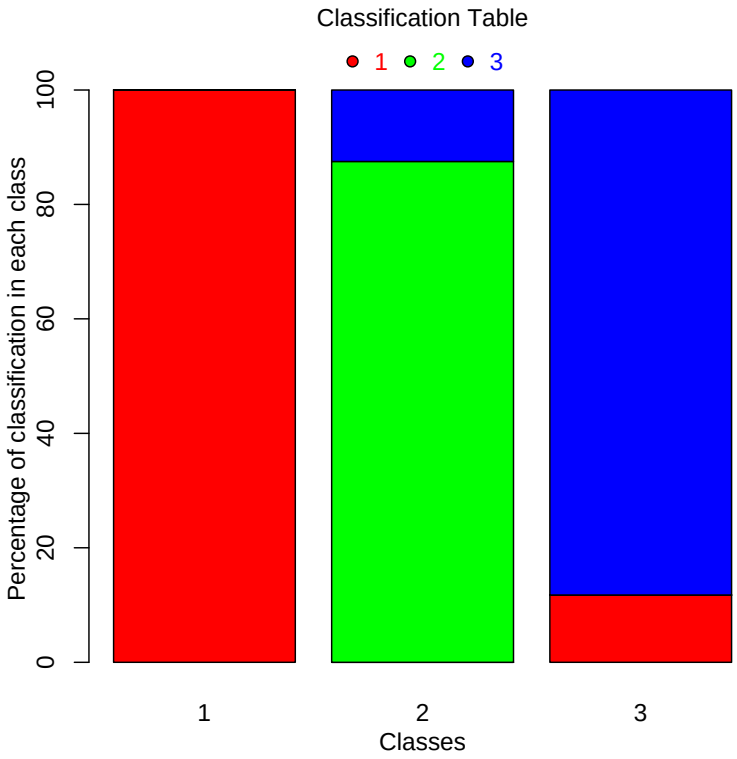

Figure 1: Plot of leave-one-out classification table for ecological data in example 1.

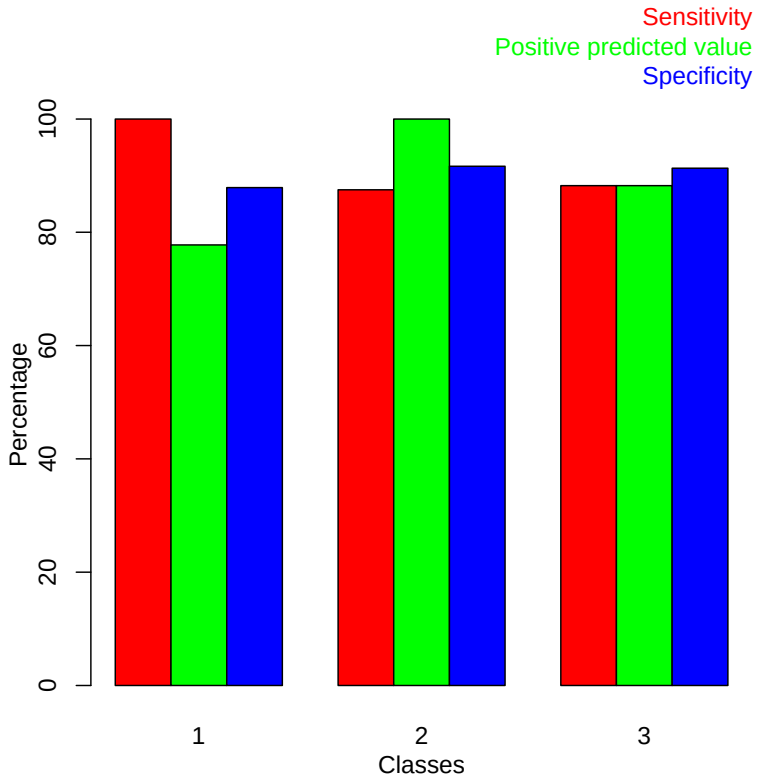

Figure 2: Plot of the sensitivity, specificity and positive predicted value for each class for ecological data in example 1.

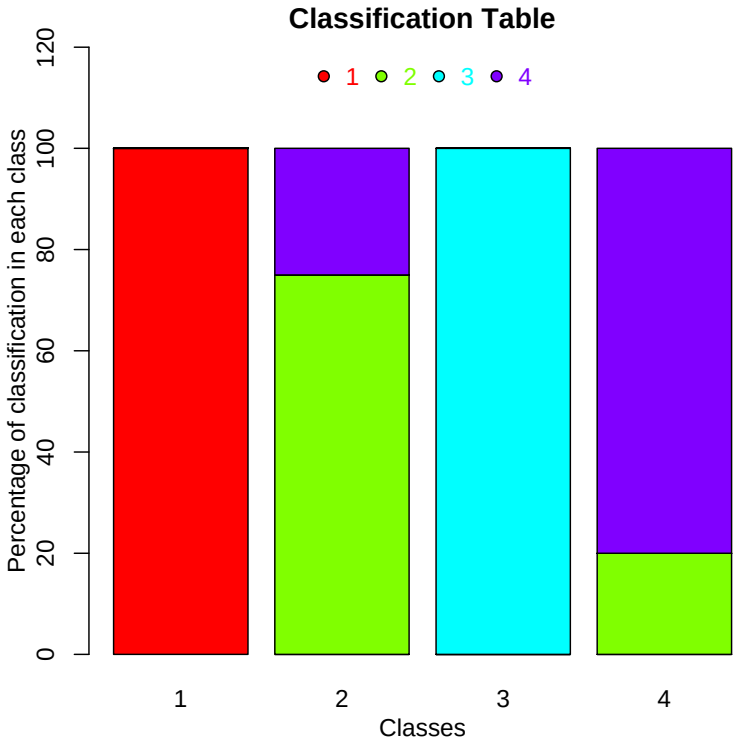

Figure 3: Plot of leave-one-out classification table for population genetics data in example 2.

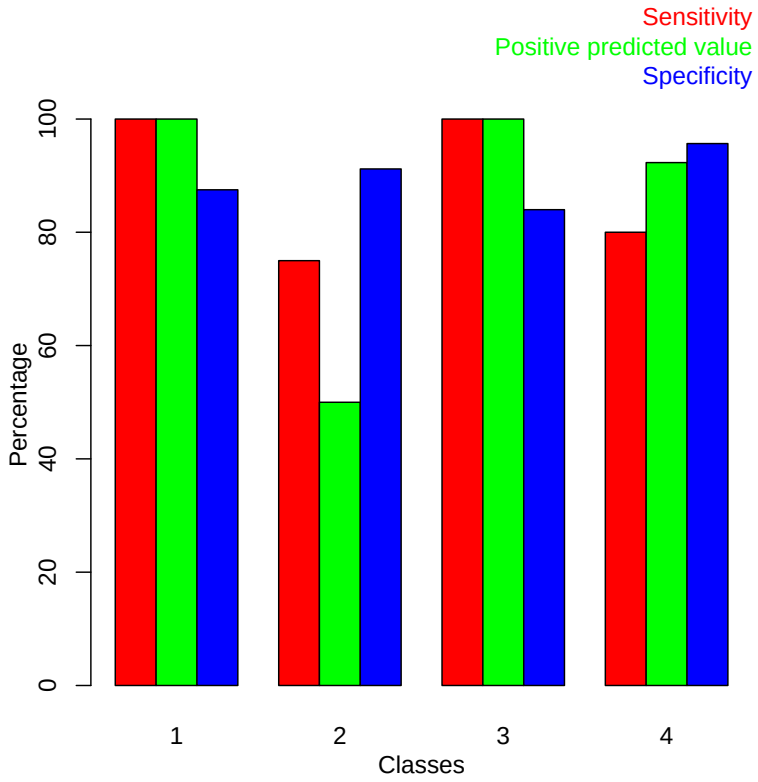

Figure 4: Plot of the sensitivity, specificity and positive predicted value for each class for population genetics data in example 2.

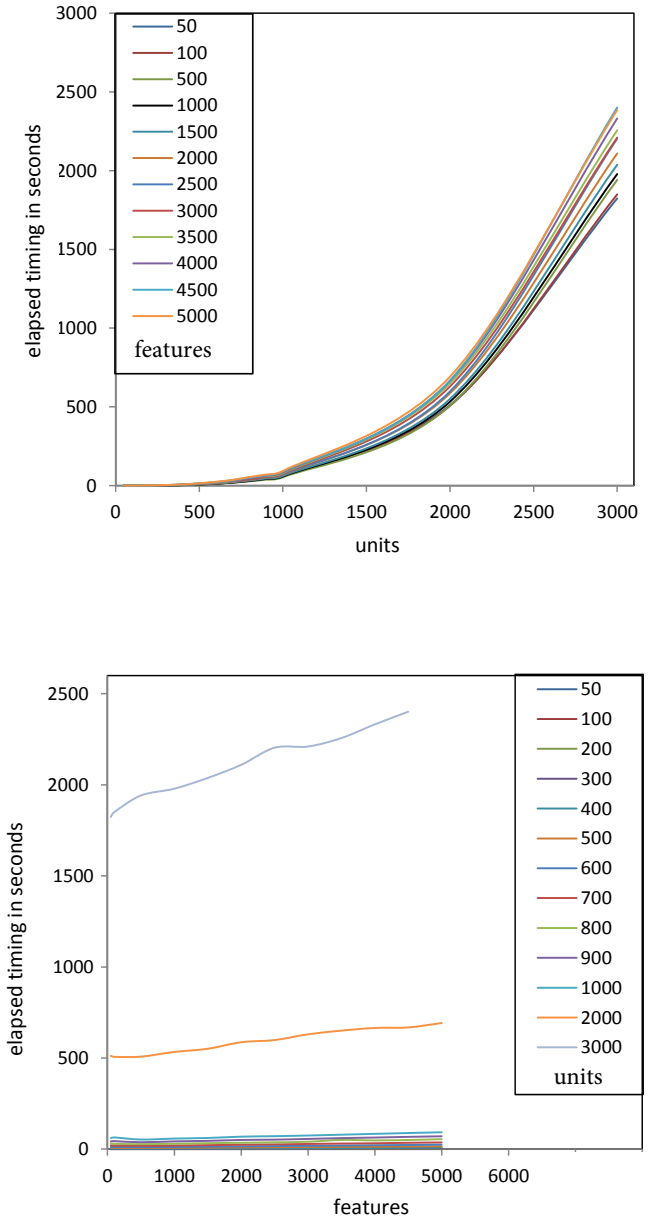

Figure 5: Artificial data sets with two classes. Top: Elapsed timing in seconds (y axes) for WDB procedure with respect to the number of units (x axes). Each line (colours in the legend) corresponds to the set with identical number of features. Bottom: Elapsed timing in seconds (y axes) for WDB procedure with respect to the number of features (x axes). Each line (colours in the legend) corresponds to the set with identical number of units.