Research Article **Open Access**

Seydou N. Sylla, Stéphane Girard*, Abdou Ka Diongue, Aldiouma Diallo, and Cheikh Sokhna

A classification method for binary predictors combining similarity measures and mixture models

DOI 10.1515/demo-2015-0017
Received June 11, 2015; accepted November 20, 2015

Abstract: In this paper, a new supervised classification method dedicated to binary predictors is proposed. Its originality is to combine a model-based classification rule with similarity measures thanks to the introduction of new family of exponential kernels. Some links are established between existing similarity measures when applied to binary predictors. A new family of measures is also introduced to unify some of the existing literature. The performance of the new classification method is illustrated on two real datasets (verbal autopsy data and handwritten digit data) using 76 similarity measures.

Keywords: Mixture model, binary predictors, kernel method, similarity measure

1 Introduction

Supervised classification aims to build a decision rule able to assign an observation x in an arbitrary space E with unknown class membership to one of L known classes $C_1, \ldots, C_L$. For building this classifier, a learning dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is used, where an observation is denoted by $x_i \in E$ and $y_i \in \{1, \ldots, L\}$ indicates the class belonging of x_i, $i = 1, \ldots, n$.

Model-based classification assumes that the predictors $\{x_1, \ldots, x_n\}$ are independent realizations of a random vector X on E and that the class conditional distribution of X is parametric. When $E = \mathbb{R}^p$, among the possible parametric distributions, the Gaussian is often preferred and, in this case, the marginal distribution of X is therefore a mixture of Gaussians. Estimation of model parameters can be achieved with maximum likelihood, see [29]. Some extensions dedicated to high-dimensional data include [6, 8, 9, 30, 31, 33, 34]. Although model-based classification is usually enjoyed for its multiple advantages, it is often limited to quantitative data. Numerous recent works focused on non Gaussian distributions such as the skew normal [43], asymmetric Laplace [16], t-distributions [1, 15] or skew t-distributions [27, 28, 45].

Only few works exist to handle categorical data using multinomial [12] or Dirichlet [5] distributions for instance. Recently, a new classification method, referred to as 'parsimonious Gaussian process Discriminant Analysis' (pgpDA), has been proposed [7] to tackle the case of data of arbitrary nature. See for instance [14] for an application to the classification of hyperspectral data. The basic idea is to introduce a kernel function in the Gaussian classification rule.

***Corresponding Author: Stéphane Girard:** Inria Grenoble Rhône-Alpes & LJK, France, E-mail: stephane.girard@inria.fr
Seydou N. Sylla: Inria Grenoble Rhône-Alpes & LJK, France
and LERSTAD-UGB, Saint-Louis, Sénégal
and URMITE-IRD, Dakar, Sénégal
Abdou Ka Diongue: LERSTAD-UGB, Saint-Louis, Sénégal
Aldiouma Diallo, Cheikh Sokhna: URMITE-IRD, Dakar, Sénégal

In this paper, we focus on the application of the pgpDA method to binary predictors. To this end, we show how new kernels can be built basing on similarity or dissimilarity measures. In particular, 76 such measures are considered. Some links are established between these measures when they are applied to binary predictors. A new family of measures is also introduced to unify the existing literature. As a result, we end up with a new supervised classification method dedicated to binary predictors combining similarity measures and mixture models. Its performance is illustrated on two real datasets (verbal autopsy data and handwritten digit data). It is shown that the proposed kernels can lead to good classification results even in challenging problems.

The paper is organized as follows. The principle of pgpDA applied to binary predictors is explained in Section 2. A brief review on similarity and dissimilarity measures is proposed in Section 3 together with some unification efforts. The construction of new kernels starting from similarity measures is presented in Section 4. The method is illustrated on real data in Section 5 and some concluding remarks are provided in Section 6. Proofs are postponed to the Appendix.

2 Classification with binary predictors using a kernel function

Conventional classification algorithms can be turned into kernel ones as far as the original method depends on the data only in terms of dot products. The dot product is simply changed to a kernel evaluation, leading to a transformation of linear algorithms to non-linear ones. Additionally, a nice property of kernel learning algorithms is the possibility to deal with any kind of data. The only condition is to be able to define a positive definite function over pairs of elements to be classified [23]. Here, we focus on binary predictors. Let us consider a learning set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $\{x_1, \ldots, x_n\}$ are assumed to be independent realizations of a random binary vector $X \in \{0, 1\}^p$. The class labels $\{y_1, \ldots, y_n\}$ are supposed to be realizations of a discrete random variable $Y \in \{1, \ldots, L\}$. They indicate the memberships of the learning data to the L classes denoted by $C_1, \ldots, C_L$, *i.e.* $y_i = k$ means that x_i belongs to the kth cluster C_k for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, L\}$.

The principle of pgpDA is as follows. Let K be a symmetric non-negative bivariate function $K : \{0, 1\}^p \times \{0, 1\}^p \to \mathbb{R}^+$. In the following, K is referred to as a kernel function and additional conditions will be assumed on K. The basic idea is to measure the proximity between individuals with K, and that close individuals are likely to belong to the same class. To this end, the kernel K computes inner products between pairs of data in some non-linear space (often referred to as a feature space). For all $k = 1, \ldots, L$, let us denote by n_k the cardinality of the class C_k, *i.e.* $n_k = \sum_{i=1}^{n} \mathbb{I}\{y_i = k\}$ where $\mathbb{I}\{.\}$ is the indicator function. We also introduce r_k the dimension of class C_k once mapped in a non-linear space with the kernel K. In practice, one has $r_k = \min(n_k, p)$ for a linear kernel and $r_k = n_k$ for the non-linear kernels considered in Section 4. See [7], Table 2 for further examples.

For all $k = 1, \ldots, L$, the function $\rho_k : \{0, 1\}^p \times \{0, 1\}^p \to \mathbb{R}^+$ is obtained by centering the kernel K with respect to the class C_k:

$$\rho_k(x, x') = K(x, x') - \frac{1}{n_k} \sum_{x_i \in C_k} (K(x_i, x') + K(x, x_i)) + \frac{1}{n_k^2} \sum_{x_i, x_j \in C_k} K(x_i, x_j). \tag{1}$$

Besides, for all $k = 1, \ldots, L$, let M_k be the $n_k \times n_k$ symmetric matrix defined by $(M_k)_{i,j} = \rho_k(x_i, x_j)/n_k$ for all $(i, j) \in \{1, \ldots, n_k\}^2$. The sorted eigenvalues of M_k are denoted by $\lambda_{k1} \geq \cdots \geq \lambda_{kn_k}$ while the associated (normed) eigenvectors are denoted by $\beta_{k1}, \ldots, \beta_{kn_k}$. In the following, β_{kji} represents the ith coordinate of β_{kj}, for $(i, j) \in \{1, \ldots, n_k\}^2$. The main assumption of the method is that the data of each class C_k live in a specific subspace of dimension d_k of the feature space (of dimension r_k). The variance of the signal in the kth group is modeled by $\lambda_{k1}, \ldots, \lambda_{kd_k}$ and the variance of the noise is modeled by λ. This amounts to supposing that the noise is homoscedastic and its variance is common to all the classes.

The classification rule introduced in [7], Proposition 2 affects $x \in \{0,1\}^p$ to the class C_ℓ if and only if $\ell = \arg\min_{k=1,\ldots,L} D_k(x)$ with

$$D_k(x) = \frac{1}{n_k}\sum_{j=1}^{d_k}\frac{1}{\lambda_{kj}}\left(\frac{1}{\lambda_{kj}}-\frac{1}{\lambda}\right)\left(\sum_{x_i\in C_k}\beta_{kji}\rho_k(x,x_i)\right)^2 + \frac{1}{\lambda}\rho_k(x,x) + \sum_{j=1}^{d_k}\log(\lambda_{kj}) + (d_{\max}-d_k)\log(\lambda) - 2\log(n_k) \tag{2}$$

where $d_{\max} = \max\{d_1,\ldots,d_L\}$ and

$$\lambda = \sum_{k=1}^{L} n_k(\text{trace}(M_k) - \sum_{j=1}^{d_k}\lambda_{kj}) \Big/ \sum_{k=1}^{L} n_k(r_k - d_k)\ .$$

Let us highlight that only the eigenvectors associated with the d_k largest eigenvalues of M_k have to be estimated. This property is a consequence of the above assumption, it allows to circumvent the unstable inversion of the matrices M_k, $k = 1,\ldots,L$ which is usually necessary in kernelized versions of Gaussian mixture models, see for instance [13, 32, 35, 44, 46]. In practice, d_k is estimated thanks to the scree-test of Cattell [11] which looks for a break in the eigenvalues scree. The selected dimension is the one for which the subsequent eigenvalues differences are smaller than a threshold t. The threshold t can be provided by the user or selected by cross-validation, see Section 5 for implementation details. The implementation of this method requires the selection of a kernel function K which measures the similarity between two binary vectors. The following invariance remark can be made:

Lemma 1. *Let K be a symmetric non-negative bivariate function $K : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$. Then, for all $\eta > 0$ and $\mu \in \mathbb{R}$, the classification rules associated with K and $\tilde{K} := \eta K + \mu$ through (2) are the same.*

As a consequence, to define a proper kernel method [23], it suffices to find a shifted version of K which is a positive definite function *i.e.*

$$\exists\, \mu \in \mathbb{R} \text{ s.t. } \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [K(x_i,x_j)+\mu] \geq 0 \text{ for all } n \in \mathbb{N},\ (c_i,c_j)\in\mathbb{R}^2,\ (x_i,x_j)\in\{0,1\}^p\times\{0,1\}^p. \tag{3}$$

The construction of kernel functions adapted to binary vectors and satisfying (3) is addressed in Section 4.

Let us highlight that pgpDA is not the only kernel-based classification method. In Section 5, pgpDA is compared to Support Vector Machine (SVM) classification [20, 21, 36] and k-nearest neighbours (kNN) [22], Chapter 13, on two real datasets. From the theoretical point of view, pgpDA offers a number of advantages compared to SVM: It is naturally a multi-class method; as a model-based classifier, it provides classification probabilities, and finally its computation cost is lower than SVM [7].

3 Similarity and dissimilarity measures

Binary similarity and dissimilarity measures play a critical role in pattern analysis problems, classification or clustering. Since the performance of these methods relies on the choice of an appropriate measure, many efforts have been made to find the most meaningful similarity measures over a hundred years, see [2, 37] for examples. The review article [37] lists 76 examples of such measures. Here, we focus on their application to binary predictors. One of the earliest measures is Jaccard's coefficient [26]. It was proposed in 1901 and it is still widely used in various fields such as ecology and biology.

Let x, x' be two vectors in $\{0,1\}^p$ and introduce $a =< x, x' >$, $b =< \mathbf{1}-x, x' >$, $c =< x, \mathbf{1}-x' >$ and $d =< \mathbf{1}-x, \mathbf{1}-x' >$, where $< .,. >$ is usual scalar product on $\mathbb{R}^p$ and $\mathbf{1} = (1,\ldots,1)^T$ in $\mathbb{R}^p$. The integer a is often referred to as the intersection of x and x', $(b+c)$ is the difference and d is the complement intersection. Note that one always has $a+b+c+d = p$.

Here, we propose to unify most of the measures proposed in the literature by introducing the following similarity measure :

$$S(x, x') = \frac{\alpha a - \theta(b + c) + \beta d}{\alpha' a + \theta'(b + c) + \beta' d} \tag{4}$$

where $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha', \beta') \in \mathbb{R}^2$ and $\theta' \neq 0$. The Symmetric Ratio Model [42] can be written as

$$S_{\text{Tversky}}(x, x') = \frac{a}{a + \theta'(b + c)}$$

and is thus a particular case of (4) where $\alpha = \alpha' = 1$ and $\theta = \beta = \beta' = 0$. Similarly, Beaulieu's similarity [3] defined by

$$S_{\text{Beaulieu}}(x, x') = \frac{-(b + c)}{\alpha' a + (b + c) + \beta' d}$$

can be obtained from (4) with $\alpha = \beta = 0$ and $\theta = \theta' = 1$. We shall also consider the particular case

$$S_{\text{Sylla \& Girard}}(x, x') = \alpha a + (1 - \alpha)d, \tag{5}$$

where $\theta = 0$, $\beta = 1 - \alpha$ and $\alpha' = \beta' = \theta' = 1/p$. This new measure can be interpreted as an extension of Intersection [37] eq. (12) and Russell & Rao [37] eq. (14) measures which both correspond to the case $\alpha = 1$. The inclusion of negative matches d in similarity measures is discussed for instance in [17, 18, 40]. It may reveal useful for instance when the classification rule depends on the coding of the data, see also Lemma 2 below. The new measure $S_{\text{Sylla \& Girard}}$ can also be seen as an extension of Sokal & Michener [37] eq. (7) and Innerproduct [37] eq. (13) measures which both correspond to the special case $\alpha = 1/2$. Thus, the parameter α in $S_{\text{Sylla \& Girard}}$ permits to balance the relative weights of positive and negative matches.
More generally, Table 1 displays 28 similarity measures from [37] which can be rewritten using our formalism (4). It appears that, on binary predictors, many similarity measures are equivalent. For instance, Hamming similarity [37] eq. (15) is equivalent to measures [37] eq. (17)–(23). Finally, some measures of [37] do not enter in our framework (4) but they can be shown to be equivalent: Forbesi measure [37] eq. (34) is equivalent to Cosine [37] eq. (31) measure, Kulczynski-II [37] eq. (41), Driver & Kroeber [37] eq. (42) and Johnson [37] eq. (43) measures are equivalent, Ochia1 measure [37] eq. (33) is equivalent to Otsuka measure [37] eq. (38), Hellinger measure [37] eq. (29) is equivalent to Chord measure [37] eq. (30) and Tarantula measure [37] eq. (75) is equivalent to Ample measure [37] eq. (76).

4 Kernels for binary predictors

The goal of this section is to build kernels adapted to binary predictors starting from the similarity and dissimilarity measures presented in Section 3. The kernels can then be plugged in the classification rule (2) to build new classification methods designed for binary predictors. In a first time, we consider the case of linear and Radial Basis Function (RBF) kernels. We then show in a second time how the RBF kernel can be extended to a wider class of exponential kernels.

Linear kernels.
Let $x, x' \in \{0, 1\}^p$. The linear kernel $K_{\text{linear}}(x, x') =< x, x' >= a$ is the simplest kernel function. In the considered binary framework, K_{linear} counts the number of positive matches between x and x'. It is shown (see [7], Proposition 3) that the associated classification rule (2) is quadratic and can thus be interpreted as a particular case of the HDDA (High Dimensional Discriminant Analysis) method [4]. Let us recall that the basic principle of HDDA is to assume that the original data of each class live in a linear subspace of low dimension. The next lemma shows that the classification rule associated with a linear kernel is independent from the coding of the data.

Table 1: Similarity measures. Measures marked with * are obtained by taking the opposite of the associated dissimilarity measures. The last column refers to the equation number in [37].

Name	α	θ	β	α'	θ'	β'	equation in [37]
Jaccard	1	0	0	1	1	0	(1)
Tanimoto	-	-	-	-	-	-	(65)
Dice	2	0	0	2	1	0	(2)
Czekanowski	-	-	-	-	-	-	(3)
Nei & li	-	-	-	-	-	-	(5)
3w-Jaccard	3	0	0	3	1	0	(4)
Sokal & Sneath-I	1	0	0	1	2	0	(6)
Sylla & Girard	α	0	$1-\alpha$	1	1	1	
Sokal & Michener	1	0	1	1	1	1	(7)
Innerproduct	-	-	-	-	-	-	(13)
Sokal & Sneath-II	2	0	2	2	1	2	(8)
Gower & Legendre	-	-	-	-	-	-	(11)
Roger & Tanimoto	1	0	1	1	2	1	(9)
Faith	1	0	0.5	1	1	1	(10)
Intersection	1	0	0	1	1	1	(12)
Russell & Rao	-	-	-	-	-	-	(14)
Hamming*	0	1	0	1	1	1	(15)
Squared-Euclid*	-	-	-	-	-	-	(17)
Canberra*	-	-	-	-	-	-	(18)
Manhattan*	-	-	-	-	-	-	(19)
Mean-Manhattan*	-	-	-	-	-	-	(20)
Cityblock*	-	-	-	-	-	-	(21)
Minkowski*	-	-	-	-	-	-	(22)
Vari*	-	-	-	-	-	-	(23)
Lance & Williams*	0	1	0	2	1	0	(27)
Bray & Curtis*	-	-	-	-	-	-	(28)
Sokal & Sneath-III	1	0	1	0	-1	0	(56)
Kulczynski-I	1	0	0	0	-1	0	(64)
Hamann	1	1	1	1	1	1	(67)

Lemma 2. *Let $x, x' \in \{0,1\}^p$ and introduce $\tilde{K}_{linear}(x,x') =< \mathbf{1}-x, \mathbf{1}-x' >= d$ (this kernel counts the number of negative matches between x and x'). Then, the classification rules (2) associated with K_{linear} and $\tilde{K}_{linear}$ are equivalent.*

Exponential kernels.

The best-known exponential kernel is RBF kernel:

$$K_{\mathrm{RBF}}(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right),$$

where σ is a positive parameter. In the binary framework, the RBF kernel can be built from the Hamming similarity measure (see Table 1 or [37] eq. (15)):

Lemma 3. *Let $x, x' \in \{0,1\}^p$. Then,*

$$K_{RBF}(x, x') = \exp\left(\frac{S_{Hamming}(x, x')}{2\sigma^2}\right).$$

We thus propose to extend this construction principle to any similarity measure S by introducing:

$$K(x, x') = \exp\left(\frac{S(x, x')}{2\sigma^2}\right). \tag{6}$$

In practice, S may be chosen to be (4), (5), or more generally in the set of 76 measures S described in [37]. The next result is the analogous of Lemma 1 for similarity measures.

Lemma 4. *Let S be a similarity measure $S : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$. Then, for all $\eta > 0$ and $\mu \in \mathbb{R}$, the classification rules associated with S and $\tilde{S} := \eta S + \mu$ through (2) and (6) are the same.*

The next result shows that any kernel defined from (4) and (6) verifies condition (3).

Proposition 1. *For all $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha', \beta') \in \mathbb{R}^2$ and $\theta' \neq 0$, the family of kernels*

$$K(x, x') = \exp\left(\frac{1}{2\sigma^2}\frac{\alpha a - \theta(b+c) + \beta d}{\alpha' a + \theta'(b+c) + \beta' d}\right)$$

defines a proper kernel classification method.

5 Experiments

The performance of the proposed method is illustrated on two real datasets described in paragraph 5.1. Some implementation details are provided in paragraph 5.2. Finally, the results are presented on paragraphs 5.3, 5.4 and 5.5.

5.1 Datasets

Verbal autopsy Data

The goal of verbal autopsy is to get some information from family about the circumstances of a death when medical certification is incomplete or absent [24]. In such a situation, verbal autopsy can be used as a routine death registration. A list of p possible symptoms is established and the collected data $X = (X_1, \ldots, X_p)$ consist of the absence or presence (encoded as 0 or 1) of each symptom on the deceased person. The probable cause of death is assigned by a physician and is encoded as a qualitative random variable Y. We refer to [39] for a review of automatic methods for assigning causes of death Y from verbal autopsy data X. In particular, classification methods based on Bayes' rule have been proposed, see [10] for instance.

Here, we focus on data measured on the deceased persons during the period from 1985 to 2010 in the three IRD (Research Institute for Development) sites (Niakhar, Bandafassi and Mlomp) in Senegal. The dataset includes $n = 2.500$ individuals (deceased persons) distributed in $L = 22$ classes (causes of death) and characterized by $p = 100$ variables (symptoms).

Binary handwritten digit data

Handwritten digit and character recognition are popular real-world tasks for testing and benchmarking classifiers, with obvious application e.g. in postal services. Here, we focus on the US Postal Service (USPS) database of handwritten digits which consists of $n = 9298$ segmented 16×16 greyscale images [25]. The dataset is available online at `http://yann.lecun.com/exdb/mnist`. The random vector X is the binarized image and

is represented as a p-dimensional vector with $p = 256$. The class to predict Y is the digit so that $L = 10$. A sample extracted from the dataset is depicted on Figure 1.

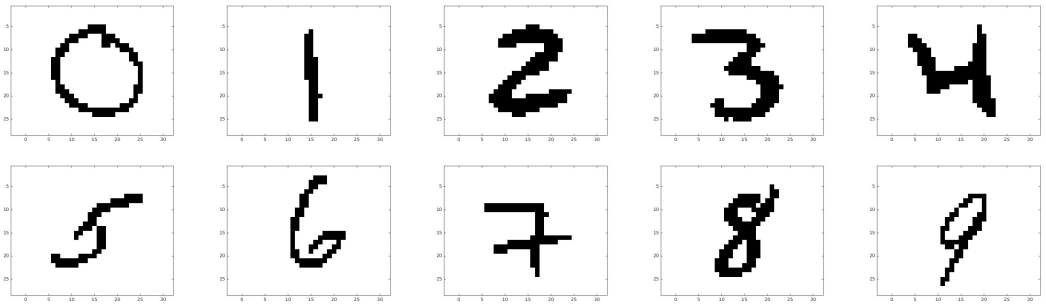

Figure 1: A sample from the binary handwritten digit data. Each pixel of a 16×16 image is either 0 (depicted in white) or 1 (depicted in black).

5.2 Experimental design

The implementation of the classification method requires the selection of the hyper-parameter $\omega = (t, \sigma)$ where t is the threshold (see Section 2) and σ is the kernel parameter see (6). To this end, a double cross-validation technique is used. The dataset of size n is randomly split $M = 50$ times into a learning set $\mathcal{L}_m$ of size τn and a test set $\mathcal{T}_m$ of size $(1 - \tau)n$ where $\tau \in (0, 1)$ is a proportion parameter and $m = 1, \ldots, M$. On each learning set $\mathcal{L}_m$, the optimal hyper-parameter $\hat{\omega}_m$ is selected by a 5-fold simple cross-validation, $m = 1, \ldots, M$. The resulting optimal hyper-parameter $\hat{\omega}$ is computed as the empirical mode of the set $\{\hat{\omega}_1, \ldots, \hat{\omega}_M\}$. Finally, the mean Correct Classification Rate (CCR) is computed on the learning sets $\mathcal{L}_m$, $m = 1, \ldots, M$ and on the test sets $\mathcal{T}_m$, $m = 1, \ldots, M$. Recall that the CCR is the percentage of well-classified observations *i.e.* the number of times that the predicted class coincides with the actual one divided by the total number of observations.

5.3 Results obtained with Sylla & Girard kernel

We first investigate the use of Sylla & Girard similarity measure (5) when plugged into (6). The CCR are computed for $\alpha \in \{0, 0.1, \ldots, 1\}$ and for several proportions τ thanks to the double cross-validation procedure described in the previous paragraph. It first appears on Figure 2 that the graphs are not symmetric with respect to $\alpha = 0.5$. This means that the coding of the observations does affect the classification. This is different from the linear case, see Lemma 2. It is also apparent that the optimal value of α does depend on the dataset. However, in both considered cases, $\alpha = 0.1$ permits to outperform the RBF kernel associated with $\alpha = 0.5$. Thus, the selection of an optimal value of α is of interest. It can be easily done by introducing α as an additional hyper-parameter in ω and thus selecting it by double cross-validation, see Paragraph 5.5 below. Finally, let us highlight that a large panel of values of α give rise to high CCR on the test set. In particular, a CCR of 87% can be reached on the challenging example of verbal autopsy data when $\tau = 78\%$ of the dataset is used to train the classifier. As a comparison, a classification based on a multinomial mixture model under conditional independence assumption yields a CCR of about 50% only [41].

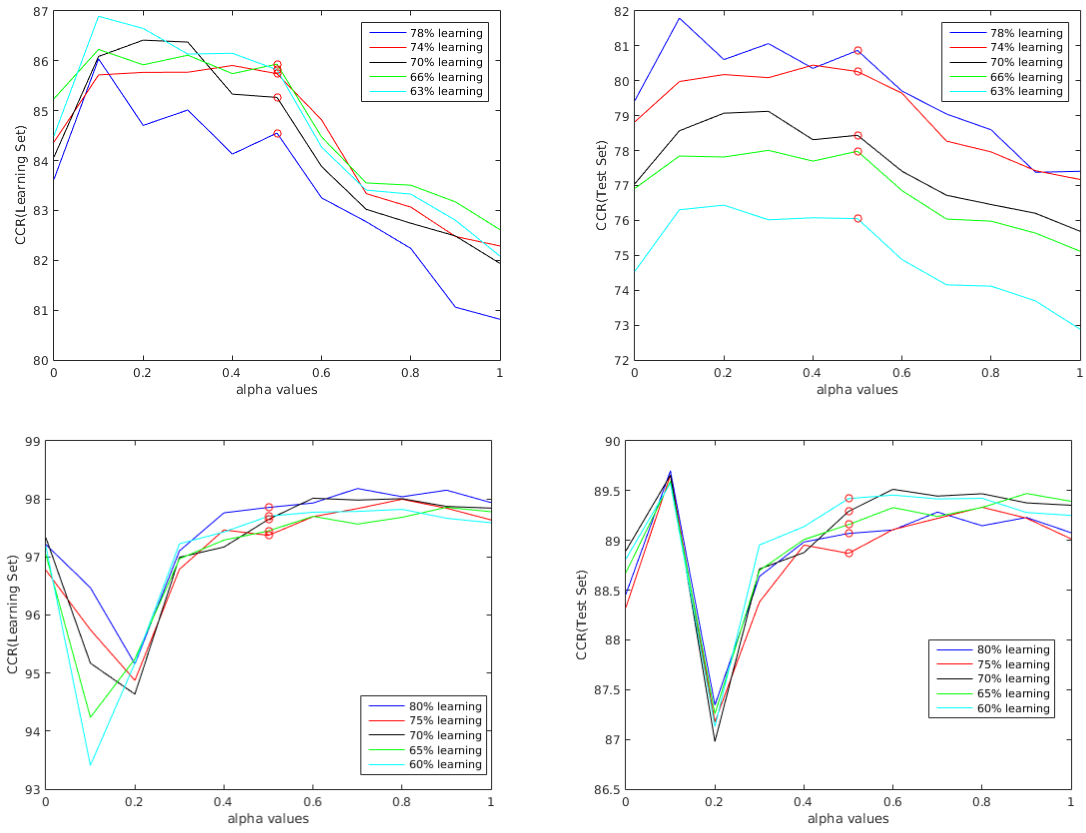

Figure 2: Correct Classification Rate (CCR) obtained with Sylla & Girard kernel (5, 6) for $\alpha \in \{0, 0.1, \ldots, 1\}$ and several proportions τ. The results obtained with the RBF kernel ($\alpha = 0.5$) are emphasized by a red circle. Left: CCR computed on the learning set, Right: CCR computed on the test set. Top: results obtained on the verbal autopsy data, bottom: results obtained on the handwritten digit data.

5.4 Results obtained with the 76 kernels from [37]

The goal of this paragraph is to compare the performance of the classification methods obtained by combining the 76 similarity and dissimilarity measures presented in [37] with the exponential kernel (6). For the sake of completeness, the results obtained with Sylla & Girard kernel presented above are also included. The classification results are summarized in Table 2 when $\tau = 63\%$ of the dataset is used to train the classifier. Only the results associated with the 18 best kernels (in terms of CCR computed on the test set) are reported. It appears that these kernels achieve good classification results on both datasets with CCR $\in [78.7\%, 89.7\%]$. It is also interesting to note that 8 kernels out of the 76 of [37] appear in the top 18 on both test datasets, namely: Euclid, Hellinger, Dice, 3w-Jaccard, Orchia1, Gower & Legendre, Roger & Tanimoto and RBF. Let us also highlight that Sylla & Girard kernel should also be included, leading to a list of 9 kernels with good results on both datasets.

5.5 Comparison with other classification methods

The proposed classification method is compared to the Random Forest method (`RandomForest` package, version 4.6-10 from `R` software), the kNN method (`fitcknn` function from the statistics and machine learning toolbox of `Matlab`) and the SVM method (library `libsvm`, version 3.2 from `Matlab`). The "one-against-all" implementation of the SVM classification method is used. SVM and Random Forest methods were used with their default parameters. In particular, in case of Random Forest method, the number of trees to grow is set

Table 2: Correct Classification Rate (CCR) on the verbal autopsy dataset (top) and on the handwritten digit dataset (bottom). The results are sorted by decreasing values of the CCR computed on the test set. The train set includes $\tau = 63\%$ individuals from the initial dataset. The last column refers to the equation number in [37].

Kernel	α	σ	threshold t	**CCR (learning set)**	**CCR (test set)**	**equation in [37]**
Euclid		4	0.60	88.0	83.8	(16)
Pearson		10	0.95	87.7	83.2	(51)
Hellinger		6	0.60	87.7	83.2	(29,30)
Dice		2	0.60	87.3	83.0	(2,3,5)
3w-Jaccard		2	0.75	87.2	82.9	(4)
Ochia1		2	0.60	87.2	82.8	(33,38)
Gower & Legendre		4	0.80	86.6	82.6	(8,11)
Roger & Tanimoto		2	0.65	85.9	82.4	(9)
Sylla & Girard	0.1	1.9	0.90	85.8	81.5	
Sylla & Girard	0.3	2.2	0.85	85.5	81.5	
Sylla & Girard = RBF	0.5	1.4	0.80	85.1	81.3	(15,17,...,23)
Sylla & Girard	0.2	1.8	0.80	85.6	81.1	
Godman & Kruskal		4	0.95	84.3	80.8	(69)
Sylla & Girard	0.4	2.5	0.80	84.7	80.6	
Sokal & Sneath 5		4	0.95	84.7	80.5	(57)
Sylla & Girard	0.6	3.09	0.80	83.2	79.6	
Sylla & Girard	0.7	3.34	0.95	83.0	79.5	
Sokal & Sneath1		2	0.05	83.4	78.7	(6)
Kernel	α	σ	threshold t	**CCR (learning set)**	**CCR (test set)**	**equation in [37]**
Hellinger		8	0.5	97.6	89.7	(29,30)
Euclid		8	0.5	97.5	89.7	(16)
Sylla & Girard	0.1	3.16	1	92.3	89.6	
Sylla & Girard	0.6	6.19	0.5	97.5	89.5	
Sylla & Girard	0.7	6.69	0.5	97.5	89.4	
Dice		2	0.5	97.4	89.4	(2,3,5)
Ochia1		2	0.5	97.4	89.4	(33,38)
Sylla & Girard = RBF	0.5	5.65	0.5	97.5	89.4	(15,17,...,23)
Roger & Tanimoto		2	0.4	97.3	89.4	(9)
Sylla & Girard	0.8	8	0.5	97.4	89.3	
Sylla & Girard	1	8	8	92.3	89.3	(9)
3w-Jaccard		4	0.5	97.3	89.3	(4)
Sylla & Girard	0.9	7.15	0.5	97.3	89.2	
Jaccard		4	0.4	97.2	89.2	(1)
Gower & Legendre		10	0.8	97.4	89.1	(8,11)
Sylla & Girard	0.4	6.3	0.5	97.2	89.1	
Sylla & Girard	0.3	5.4	0.5	96.9	88.7	
Sylla & Girard	0.2	4.4	0.45	97.9	86.8	

to `ntree=500` and the minimum size of terminal nodes is set to `nodesize=1`. Some additional experiments reported in Table 5 and Table 6 showed that the obtained classifications were not very sensitive to these parameters: The CCR computed on the test set remains approximately constant when `nodesize` $\in \{1, \ldots, 10\}$ and `ntree` $\in \{250, 500, 750, 1000\}$. The number k of neighbours in kNN method is selected using the dou-

ble cross-validation procedure. Sylla & Girard kernel is plugged into pgpDA, kNN and SVM methods with $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. The selection of α by double cross-validation has also been implemented, the resulting value is denoted by $\alpha^\star$ in the following.

Table 3: Correct Classification Rate (CCR) on the verbal autopsy dataset (top) and on the handwritten digit dataset (bottom). Sylla & Girard kernel is plugged into pgpDA, SVM and kNN methods for $\alpha \in \{0.1, 0.2, \ldots, 1\}$. The CCR associated with the parameter $\alpha^\star$ selected by the double cross-validation procedure is emphasized. The train set includes $\tau = 63\%$ individuals from the initial dataset.

	pgpDA		**SVM**		**kNN**	
α	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**
0.1	86.9	76.3	85.3	74.6	64.5	53.1
0.2	86.6	76.4	**79.9**	**70.8**	67.4	57.6
0.3	**86.1**	**76.0**	79.5	70.4	**68.3**	**59.5**
0.4	86.1	76.1	76.0	67.9	69.1	60.9
0.5	85.8	76.1	72.7	65.3	69.0	61.0
0.6	84.3	74.9	70.3	63.5	69.2	61.8
0.7	83.4	74.2	69.2	62.6	68.3	60.9
0.8	83.3	74.1	68.7	62.2	68.5	60.9
0.9	82.8	73.7	68.2	61.7	67.7	59.8
1	82.1	72.0	67.6	61.2	64.6	56.4
	pgpDA		**SVM**		**kNN**	
α	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**
0.1	93.4	89.6	**100.0**	**93.1**	91.5	91.4
0.2	95.2	87.1	99.9	97.5	94.3	93.8
0.3	97.2	88.9	99.9	97.8	**95.5**	**94.3**
0.4	**97.4**	**89.1**	99.7	97.7	95.3	93.4
0.5	97.7	89.4	99.4	97.4	94.7	92.0
0.6	97.8	89.4	99.3	97.2	92.5	88.7
0.7	97.8	89.4	99.1	97.0	89.3	83.5
0.8	97.8	89.4	98.3	96.2	82.5	74.7
0.9	97.7	89.3	98.0	96.0	72.5	62.2
1	97.6	89.3	97.7	95.7	56.1	45.2

Table 4: Correct Classification Rate (CCR) obtained with Random Forest (`nodesize=1` and `ntree=500`). The training set includes $\tau = 63\%$ individuals from the initial dataset.

	Random Forest	
	CCR (learning set)	**CCR (test set)**
Verbal autopsy	88.7	67.4
Handwritten digit	100.0	94.0

Table 5: Correct Classification Rate (CCR) obtained with Random Forest for several values of `nodesize` and `ntree` on the verbal autopsy dataset. The CCR obtained with the default parameters `nodesize=1` and `ntree=500` are emphasized, and reported in Table 4. The training set includes $\tau = 63\%$ individuals from the initial dataset.

					CCR (training set)					
nodesize	1	2	3	4	5	6	7	8	9	10
ntree										
250	88.8	87.9	86.9	85.5	83.9	82.7	81.4	79.9	78.1	76.6
500	**88.7**	88.0	86.9	85.6	84.4	82.9	81.6	79.9	78.3	76.7
750	88.8	88.2	87.0	85.7	84.4	82.9	81.5	79.6	78.1	76.6
1000	88.7	88.2	87.2	85.7	84.2	83.1	81.6	80.0	78.1	76.8
					CCR (test set)					
nodesize	1	2	3	4	5	6	7	8	9	10
ntree										
250	67.3	67.1	67.0	67.6	67.0	67.1	66.8	66.3	66.1	65.9
500	**67.4**	67.8	67.3	67.0	67.4	66.9	66.7	66.4	65.9	65.7
750	67.5	67.6	67.4	67.1	67.2	66.9	66.7	66.5	65.8	65.8
1000	67.9	67.7	67.3	67.3	67.2	67.0	66.7	66.6	66.2	65.5

Table 6: Correct Classification Rate (CCR) obtained with Random Forest for several values of `nodesize` and `ntree` on the handwritten digit dataset. The CCR obtained with the default parameters `nodesize=1` and `ntree=500` are emphasized, and reported in Table 4. The training set includes $\tau = 63\%$ individuals from the initial dataset.

					CCR (training set)					
nodesize	1	2	3	4	5	6	7	8	9	10
ntree										
250	100.0	100.0	100.0	99.9	99.9	99.8	99.6	99.4	99.2	98.9
500	**100.0**	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.2	99.0
750	100.0	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.3	99.0
1000	100.0	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.3	99.0
					CCR (test set)					
nodesize	1	2	3	4	5	6	7	8	9	10
ntree										
250	93.9	93.8	93.8	93.6	93.5	93.5	93.2	93.1	93.1	93.0
500	**94.0**	94.0	93.7	93.8	93.7	93.5	93.3	93.3	93.1	93.2
750	93.9	94.0	93.8	93.8	93.7	93.5	93.5	93.3	93.3	93.1
1000	94.0	94.0	93.9	93.8	93.7	93.6	93.4	93.4	93.3	93.2

It appears in Table 3 and Table 4 that, on the verbal autopsy dataset, pgpDA method yields better results than SVM, kNN and Random Forest methods on the test set. Since, on the learning set, the CCR obtained with Random Forest is larger than the CCR associated with pgpDA, kNN and SVM methods for all values of α, one can suspect that Random Forest overfits this dataset. One can also observe that the CCR associated with pgpDA slightly depends on α (CCR $\in [72.0\%, 76.4\%]$) whereas CCR associated with SVM and kNN are very sensitive to α (CCR $\in [61.2\%, 74.6\%]$ and CCR $\in [53.1\%, 61.8\%]$ respectively). At the opposite, SVM, kNN and Random Forest yield better results than pgpDA on the handwritten digit dataset. The CCR associated with pgpDA is however satisfying, it is larger than 87.1% whatever the value of α is. This may due to the small number of classes ($L = 10$ here, $L = 22$ in the previous situation) which makes the classification problem not so difficult.

The selection by double cross-validation of the parameter α in Sylla & Girard achieves good results for all the considered classification methods. The selected value remains stable accross the experiments: $\alpha^\star \in \{0.3, 0.4\}$ with pgpDA, $\alpha^\star \in \{0.1, 0.2\}$ with SVM and $\alpha^\star = 0.3$ for kNN. It is a first step towards an automatic choice of the similarity measure in the classification framework. Finally, let us precise that the experiments were conducted on a two processor computer (8 cores cadenced a 2.6 GHz). The computations on one learning set $\mathcal{L}_m$ from the handwritten digit dataset took respectively 35 minutes (pgpDA), 40 minutes (SVM), 48 minutes (kNN) and 11 minutes (Random Forest).

6 Conclusion

This work was motivated by two facts: First, numerous binary similarity measures have been used in various scientific fields. Second, model-based mixtures offer a coherent response to the problem of classification by providing classification probabilities and natural multi-class support. Basing on these remarks, our main contribution is the proposal of a new classification method combining mixture models and binary similarity measures. The method provides good classification performances on challenging data sets (high number of variables and classes). We believe that this method can reveal useful in a wide variety of classification problems with binary predictors. As a by-product of this work, some new similarity measures are proposed to unify the existing literature.

This work could be extended to the classification of mixed quantitative and binary predictors. As suggested in [7], to deal with such data, one can build a combined kernel by mixing a kernel based on a similarity measure (as proposed here) for the binary predictors and a RBF kernel for the quantitative ones. The combined kernel could be for instance the weighted sum or the product of the two kernels, see [19] for further details on multiple kernel learning.

Appendix: Proofs

Proof of Lemma 1.
For all $k = 1, \dots, L$, let $\tilde{\rho}_k$ be the function defined similarly to (1) by

$$\begin{aligned}
\tilde{\rho}_k(x, x') &:= \tilde{K}(x, x') - \frac{1}{n_k} \sum_{x_i \in C_k} (\tilde{K}(x_i, x') + \tilde{K}(x, x_i)) + \frac{1}{n_k^2} \sum_{x_i, x_j \in C_k} \tilde{K}(x_i, x_j) \\
&= \eta K(x, x') - \frac{1}{n_k} \sum_{x_i \in C_k} (\eta K(x_i, x') + \eta K(x, x_i)) + \frac{1}{n_k^2} \sum_{x_i, x_j \in C_k} \eta K(x_i, x_j), \\
&= \eta \rho_k(x, x').
\end{aligned}$$

Thus, $(\tilde{M}_k)_{i,j} := \tilde{\rho}_k(x_i, x_j)/n_k = \eta (M_k)_{i,j}$ for all $(i, j) \in \{1, \dots, n_k\}^2$. Let the sorted eigenvalues of $\tilde{M}_k$ be denoted by $\tilde{\lambda}_{k1} \geq \cdots \geq \tilde{\lambda}_{kn_k}$ and the associated (normed) eigenvectors be denoted by $\tilde{\beta}_{k1}, \dots, \tilde{\beta}_{kn_k}$. Clearly, $\tilde{\lambda}_{kj} = \eta \lambda_{kj}$ and $\tilde{\beta}_{kj} = \pm \beta_{kj}$ for all $(j, k) \in \{1, \dots, n_k\}^2$. It follows that

$$\tilde{\lambda} := \sum_{k=1}^{L} n_k (\mathrm{trace}(\tilde{M}_k) - \sum_{j=1}^{d_k} \tilde{\lambda}_{kj}) \Big/ \sum_{k=1}^{L} n_k (r_k - d_k) = \eta \lambda$$

and therefore

$$\begin{aligned}\tilde{D}_k(x) &:= \frac{1}{n_k}\sum_{j=1}^{d_k}\frac{1}{\tilde{\lambda}_{kj}}\left(\frac{1}{\tilde{\lambda}_{kj}}-\frac{1}{\tilde{\lambda}}\right)\left(\sum_{x_i\in C_k}\tilde{\beta}_{kji}\tilde{\rho}_k(x,x_i)\right)^2+\frac{1}{\tilde{\lambda}}\tilde{\rho}_k(x,x)\\ &+ \sum_{j=1}^{d_k}\log(\tilde{\lambda}_{kj})+(d_{\max}-d_k)\log(\tilde{\lambda})-2\log(n_k)\\ &= D_k(x)+d_{\max}\log\eta.\end{aligned}$$

Since $d_{\max}\log\eta$ does not depend on k, the two classification rules are equivalent. □

Proof of Lemma 2.
To simplify the notations, let $K(x,x') :=< x,x' >$ and

$$\begin{aligned}\tilde{K}(x,x') &:= <\mathbf{1}-x,\mathbf{1}-x'>\\ &= <\mathbf{1},\mathbf{1}>-<\mathbf{1},x>-<\mathbf{1},x'>+<x,x'>\\ &= K(1,1)-K(\mathbf{1},x)-K(\mathbf{1},x')+K(x,x').\end{aligned}$$

For all $k=1,\ldots,L$, replacing in

$$\tilde{\rho}_k(x,x') := \tilde{K}(x,x')-\frac{1}{n_k}\sum_{x_i\in C_k}(\tilde{K}(x_i,x')+\tilde{K}(x,x_i))+\frac{1}{n_k^2}\sum_{x_i,x_j\in C_k}\tilde{K}(x_i,x_j),$$

yields $\tilde{\rho}_k(x,x')=\rho_k(x,x')$ in view of (1) and thus the two classification rules are equivalent. □

Proof of Lemma 3.
For all $x,x'\in\{0,1\}^p$, we have

$$\begin{aligned}\|x-x'\|^2 &= \sum_{i=1}^p x_i^2+\sum_{i=1}^p(x_i')^2-2\sum_{i=1}^p x_ix_i'\\ &= \sum_{i=1}^p x_i+\sum_{i=1}^p x_i'-2\sum_{i=1}^p x_ix_i'\\ &= \sum_{i=1}^p x_i(1-x_i')+\sum_{i=1}^p x_i'(1-x_i)\\ &= b+c,\end{aligned}$$

and the conclusion follows. □

Proof of Lemma 4.
Let us remark that

$$\tilde{K}(x,x') := \exp\left(\frac{\tilde{S}(x,x')}{2\sigma^2}\right)=\exp\left(\frac{\eta S(x,x')+\mu}{2\sigma^2}\right)=\eta'\exp\left(\frac{S(x,x')}{2\sigma'^2}\right)$$

with $\eta'=\exp(\mu/(2\sigma^2))$ and $\sigma'=\sigma/\sqrt{\eta}$. The conclusion follows from Lemma 1. □

Proof of Proposition 1.
Let us introduce

$$\begin{aligned} S_1(x, x') &:= \alpha a - \theta(b + c) + \beta d, \\ S_2(x, x') &:= \alpha' a + \theta'(b + c) + \beta' d, \end{aligned}$$

such that

$$K(x, x') = \exp\left(\frac{1}{2\sigma^2}\frac{S_1(x, x')}{S_2(x, x')}\right).$$

– Let us first prove that S_1 defines a proper kernel classification method. Note that, if $\theta = 0$, then $S_1(x, x') = \alpha K_{\text{linear}}(x, x') + \beta \tilde{K}_{\text{linear}}(x, x')$ and the conclusion follows. In the case where $\theta > 0$, one can write

$$S_1(x, x') = \alpha a - \theta(p - a - d) + \beta d = \theta p(ua + vd - 1)$$

with $u := (1 + \alpha/\theta)/p > 0$ and $v := (1 + \beta/\theta)/p > 0$. It is thus clear that S_1 verifies condition (3).

– The second step consists in showing that $1/S_2$ defines a proper kernel classification method. Let us focus on the case where $0 \leq \alpha', \beta' < \theta'$, the other cases being similar. Introduce $u' := (1 - \alpha'/\theta')/p > 0$ and $v' := (1 - \beta'/\theta')/p > 0$ such that

$$S_2(x, x') = \alpha' a + \theta'(p - a - d) + \beta' d = \theta' p[1 - (u' a + v' d)]$$

with $u' \in [0, 1)$ and $v' \in [0, 1)$. Since $0 \leq u' a + v' d < 1$, the following expansion holds:

$$\frac{1}{S_2(x, x')} = \frac{1}{\theta' p}\sum_{i=0}^{\infty}(u' a + v' d)^i.$$

For all $N > 0$, let

$$S_{3,N}(x, x') := \frac{1}{\theta' p}\sum_{i=0}^{N}(u' a + v' d)^i.$$

Since $S_{3,N}$ is obtained from sums and products of K_{linear} and $\tilde{K}_{\text{linear}}$, it follows from [38], Proposition 3.22 (i) and (iii) that $S_{3,N}$ defines a proper kernel classification method for all $N > 0$. As a consequence, $S_{3,N}$ verifies condition (3) for all $N > 0$. Letting $N \to \infty$, one can conclude that $1/S_2$ defines a proper kernel classification method.

– Finally, in view of [38], Proposition 3.22 (ii), (iii) and Proposition 3.24 (ii), it follows that K defines a proper kernel classification method. □

Acknowledgement: The authors would like to greatly thank the referees for their helpful remarks and comments on the manuscript.

References

[1] Andrews, J.L. and P.D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Stat. Comp.* 22(5), 1021–1029.

[2] Batagelj, V. and M. Bren (1995). Comparing resemblance measures. *J. Classif.* 12, 73–90.

[3] Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *J. Classif.* 6, 233–246.

[4] Bergé, L., C. Bouveyron, and S. Girard. (2012). HDclassif: an R package for model-based clustering and discriminant analysis of high-dimensional data. *J. Stat. Softw.* 46(6), 1–29.

[5] Bouguila, N., D. Ziou, and J. Vaillancourt (2003). Novel mixtures based on the Dirichlet distribution: application to data and image classification. In *Machine Learning and Data Mining in Pattern Recognition*, Perner P. ed., 172–181, Springer-Verlag, Berlin Heidelberg.

[6] Bouveyron, C. and C. Brunet (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Stat. Comp.* 22, 301–324.

[7] Bouveyron, C., M. Fauvel and S. Girard (2015). Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Stat. Comp.*, 25, 1143–1162.
[8] Bouveyron, C., S. Girard and C. Schmid (2007). High-dimensional discriminant analysis. *Commun. Stat. A-Theor.* 36, 2607–2623.
[9] Bouveyron, C., S. Girard and C. Schmid (2007). High-dimensional data clustering. *Comput. Stat. Data An.* 52, 502–519.
[10] Byass, P., D.L. Huong and H.V. Minh (2003). A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam. *Scand. J. Public Health* 31(62), 32–37.
[11] Cattell, R. (1966). The scree test for the number of factors. *Multivar. Behav. Res.* 1(2), 245–276.
[12] Celeux, G. and G. Govaert (1991). Clustering criteria for discrete data and latent class models. *J. Classif.* 8, 157–176.
[13] Dundar, M.M. and D.A. Landgrebe (2004). Toward an optimal supervised classifier for the analysis of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* 42(1), 271–277.
[14] Fauvel, M., C. Bouveyron and S. Girard (2015). Parsimonious Gaussian process models for the classification of hyperspectral remote sensing images. *IEEE Geosci. Remote Sens. Lett.*, to appear.
[15] Forbes, F. and D. Wraith (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tail-weight: application to robust clustering. *Stat. Comp.* 24(6), 971–984.
[16] Franczak, B.C., R.P. Browne and P.D. McNicholas (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(6), 1149–1157.
[17] Goodman, L.A and W.H. Kruskal (1954). Measures of association for cross classifications. *J. Amer. Statist. Assoc.* 49, 732–764.
[18] Goodman, L.A and W.H. Kruskal (1959). Measures of association for cross classifications II. Further discussion and references. *J. Amer. Statist. Assoc.* 54, 35–75.
[19] Gönen, M. and E. Alpaydin (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.
[20] Guermeur, Y. (2002). Combining discriminant models with new multi-class SVMs. *Pattern Anal. Appl.* 5(2), 168–179.
[21] Guermeur, Y. (2007). VC theory of large margin multi-category classifiers. *J. Mach. Learn. Res.* 8, 2551–2594.
[22] Hastie, T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning*. Second edition. Springer, Berlin.
[23] Hofmann, T., B. Schölkopf and A. Smola (2008). Kernel methods in machine learning. *Annals Stat.* 36(3), 1171–1220.
[24] Huong, D.L., H.V. Minh and P. Byass (2003). Applying verbal autopsy to determine cause of death in rural Vietnam. *Scand. J. Public Health* 31(62), 19–25.
[25] LeCun, Y., L. Bottou, Y. Bengio and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of IEEE* 86(11), 2278–2324.
[26] Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.
[27] Lee, S. and G. McLachlan (2013). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Stat. Comp.* 24(2), 181–202.
[28] Lin, T.I. (2010). Robust mixture modeling using multivariate skew t-distribution. *Stat. Comp.* 20, 343–356.
[29] McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
[30] McLachlan, G., D. Peel and R. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data An.* 41, 379–388.
[31] McNicholas, P. and B. Murphy (2008). Parsimonious Gaussian mixture models. *Stat. Comp.* 18, 285–296.
[32] Mika, S., G. Ratsch, J. Weston, B. Schölkopf and K.R. Müller (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson and S. Douglas eds., 41–48. The Institute of Electrical and Electronics Engineers, Inc. New York.
[33] Montanari, A. and C. Viroli (2010). Heteroscedastic factor mixture analysis. *Stat. Modeling* 10, 441–460.
[34] Murphy, T.B., N. Dean and A.E. Raftery (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Annals Appl. Stat.* 4, 219–223.
[35] Pekalska, E. and B. Haasdonk (2009). Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(6), 1017–1032.
[36] Scholkopf, B. and A.J. Smola (1990). *Learning with Kernels*. The MIT Press, Cambridge MA.
[37] Seung-Seok, C., C. Sung-Hyuk and C. Tappert (2010). A survey of binary similarity and distance measures. *J. Syst. Cybern. Informatics* 8, 43–48.
[38] Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press.
[39] Reeves, B.C. and M.A. Quigley (1997). A review of data-derived methods for assigning causes of death from verbal autopsy data. *Int. J. Epidemiol.* 26, 1080–1089.
[40] Sneath, P.H.A. and R.R. Sokal (1973). *Numerical Taxonomy: the Principles and Practice of Numerical Classification*, W.H. Freeman and Company, San Francisco.
[41] Sylla, S., S. Girard, A. Diongue, A. Diallo and C. Sokhna (2014). Classification supervisée par modèle de mélange: Application aux diagnostics par autopsie verbale. *46èmes Journées de Statistique organisées par la Société Française de Statistique*, Rennes.
[42] Tversky, A. (1977). Feature of similarity, *Psychol. Rev.* 84, 327–352.

[43] Vilca, F., N. Balakrishnan and C. Zeller (2014). Multivariate skew-normal generalized hyperbolic distribution and its properties. *J. Multivar. Anal.* 128, 73–85.

[44] Wang, J., J. Lee and C. Zhang (2003). Kernel trick embedded Gaussian mixture model. In *Algorithmic Learning Theory*, Gavalda, R., Jantke, K. P., Takimoto, E. eds., 159–174. Springer-Verlag, Berlin Heidelberg.

[45] Wraith, D. and F. Forbes (2015). Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering. *Comput. Stat. Data An.* 90, 61–73.

[46] Xu, Z., K. Huang, J. Zhu, I. King and M.R. Lyu (2009). A novel kernel-based maximum a posteriori classification method. *Neural Networks* 22, 977–987, 2009.