A supervised binary model-based classification method combining similarity measures and mixture models

Seydou Nourou Sylla, Stéphane Girard, Abdou Ka Diongue, Aldiouma Diallo, Cheikh Sokhna

A supervised binary model-based classification method combining similarity measures and mixture models

Seydou N. SYLLA [1,2,3], Stéphane GIRARD [1,⋆], Abdou Ka DIONGUE [2]
Aldiouma DIALLO [3] & Cheikh SOKHNA [3]

[1] *Inria Grenoble Rhône-Alpes & LJK, France,*
[2] *LERSTAD-UGB, Saint-Louis, Sénégal,*
[3] *URMITE-IRD, Dakar, Sénégal,*
[⋆] *Corresponding author,* stephane.girard@inria.fr

Abstract. In this paper, a new supervised classification method dedicated to binary data 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 data. 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 hand-digit data) using 76 similarity measures.

Keywords. Mixture model, binary data, 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 $C_1, ..., C_L$. For building this classifier, a learning dataset $\{(x_1, y_1), ..., (x_n, y_n)\}$ is used, where an observation is denoted by $x_\ell \in E$ and $y_\ell \in \{1, ..., k\}$ indicates the class belonging of x_ℓ, $\ell = 1, \ldots, n$.

Model-based classification assumes that $\{x_1, ..., x_n\}$ 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 [26]. Some extensions dedicated to high-dimensional data include [5, 7, 8, 27, 28, 30, 31]. 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 [40], asymmetric Laplace [15], t-distributions [1, 13] or skew t-distributions [14, 24, 25].

Only few works exist to handle categorical data using multinomial [11] or Dirichlet [4] distributions for instance. Recently, a new classification method, referred to as pgpDA, has been

proposed [6] to tackle the case of data of arbitrary nature. The basic idea is to introduce a kernel function in the Gaussian classification rule.

In this paper, we focus on the application of the ppgDA method to binary data. 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 data. 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 data combining similarity measures and mixture models. Its performance is illustrated on two real datasets (verbal autopsy data and hand-digit data).

The paper is organized as follows. The principle of pgpDA applied to binary data is explained in Section 2. A brief review on similarity and dissimilarity measures in 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 Binary classification using a kernel function

Conventional classification algorithms can be turned into kernel ones as soon 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 [20]. Here, we focus on binary data. 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 assumed to be realizations of a discrete random variable $Y \in \{1, \ldots, L\}$. It indicates 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. 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:

$$\rho_k(x, x') = K(x, x') - \frac{1}{n_k} \sum_{x_\ell \in C_k} (K(x_\ell, x') + K(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} K(x_\ell, x_{\ell'}),$$

where n_k is the cardinality of the class C_k, *i.e.* $n_k = \sum_{i=1}^{n} \mathbb{I}\{y_i = k\}$ and with $\mathbb{I}\{.\}$ the indicator function. Besides, for all $k = 1, \ldots, L$, let us introduce the $n_k \times n_k$ symmetric matrix M_k defined by $(M_k)_{\ell,\ell'} = \rho_k(x_\ell, x_{\ell'})/n_k$ for all $(\ell, \ell') \in \{1, \ldots, n_k\}^2$. 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, $\beta_{kj\ell}$ represents the ℓth cordinate of β_{kj}, for $(j, \ell) \in \{1, \ldots, n_k\}^2$. The classification rule introduced in [6], Proposition 2 affects $x \in \{0,1\}^p$ to the class C_i if and only if $i = \arg\min_{k=1,\ldots,L} D_k(x)$ with

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

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

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

Here, r_k is the dimension of class C_k once mapped in a nonlinear 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 nonlinear kernels considered in Section 4. See [6], Table 2 for further examples. Moreover, 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 crucial assumption of this method: The data of each class C_k live in a specific subspace (of dimension d_k) of the space (of dimension r_k) defined by the kernel K. This assumption 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 [12, 29, 32, 41, 42]. In practice, d_k is estimated thanks to the scree-test of Cattell [10] 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 (1) are the same.*

As a consequence, to define a proper kernel method [20], it suffices to find a shifted version of

K which is a positive definite function *i.e.*

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

The construction of kernel functions adapted to binary vectors and satisfying (2) 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 [18, 19, 33] 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 [6].

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, 34] for examples. The review article [34] lists 76 examples of such measures. Here, we focus on their application to binary data. One of the earliest measure is Jaccard's coefficient [23]. It was proposed in 1901 and it is still widely used in the various fields such as ecology and biology. We also refer to [16, 17, 37] for a discussion on the inclusion or exclusion of negative matches in similarity measures.

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 :

$$S(x, x') = \frac{\alpha a - \theta(b+c) + \beta d}{\alpha' a + \theta'(b+c) + \beta' d} \tag{3}$$

where $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha', \beta') \in \mathbb{R}^2$ and $\theta' \neq 0$. The Symmetric Ratio Model [39] can be written as

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

and is thus a particular case of (3) 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 (3) with $\alpha = \beta = 0$ and $\theta = \theta' = 1$. We shall also consider the particular case

$$S_{\text{Sylla \& Girard}}(x, x') = \alpha a + (1-\alpha)d, \tag{4}$$

where $\theta = 0$, $\beta = 1 - \alpha$ and $\alpha' = \beta' = \theta' = 1/p$. Sokal & Michener (7) and Innerproduct (13) measures in [34] are special cases of $S_{\text{Sylla \& Girard}}$ obtained with $\alpha = 1/2$ while Intersection (12) and Russell & Rao (14) measures correspond to the case $\alpha = 1$.
More generally, Table 1 shows 28 similarity measures from [34] which can be rewritten using our formalism (3). It appears that, on binary data, many different similarity measures are equivalent. For instance, Hamming similarity (15) is equivalent to measures (17)–(23). Finally, some measures of [34] do not enter in our framework (3) but they can be shown to be equivalent: Forbesi measure (34) is equivalent to Cosine (31) measure, Kulczynski-II (41), Driver & Kroeber (42) and Johnson (43) measures are equivalent, Ochia1 measure (33) is equivalent to Otsuka measure (38), Hellinger measure (29) is equivalent to Chord measure (30) and Tarantula measure (75) is equivalent to Ample measure (76).

4 Kernels for binary data

The goal of this section is to build kernels adapted to binary data starting from the similarity and dissimilarity measures presented in Section 3. The kernels can then be plugged in the classification rule (1) to build new classification methods designed for binary data. 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 [6], Proposition 3) that the associated classification rule (1) is quadratic and can thus be interpreted as a particular case of the HDDA method [7]. The next lemma shows that the classification rule associated with a linear kernel is independent from the coding of the data.

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 (1) associated with K_{linear} and $\tilde{K}_{linear}$ are equivalent.*

Name	α	θ	β	α'	θ'	β'	**equation**
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)

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 [34].

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, this kernel can be rewritten as

$$K_{\mathrm{RBF}}(x, x') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right) = \exp\left(-\frac{(b+c)}{2\sigma^2}\right) = \exp\left(\frac{S_{\mathrm{Hamming}}(x_\ell, x_{\ell'})}{2\sigma^2}\right).$$

It appears that the RBF kernel can be built from the Hamming similarity measure (see Table 1 or [34], equation (15)). We thus propose to extend this construction principle to any similarity measure S by introducing:

$$K(x, x') = \exp\left(\frac{S(x, x')}{2\sigma^2}\right). \tag{5}$$

In practice, S may be chosen to be (3), (4), or more generally in the set of 76 measures S described in [34]. The next result is the analogous of Lemma 1 for similarity measures.

Lemma 3. *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 (1) and (5) are the same.*

The next result shows that any kernel defined from (5) and (3) verifies condition (2).

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 [21]. 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 [36] 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 [9] 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 [22]. The dataset is avalaible online at `http://yann.lecun.com/exdb/mnist`. The random vector X is the binarized image and 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.

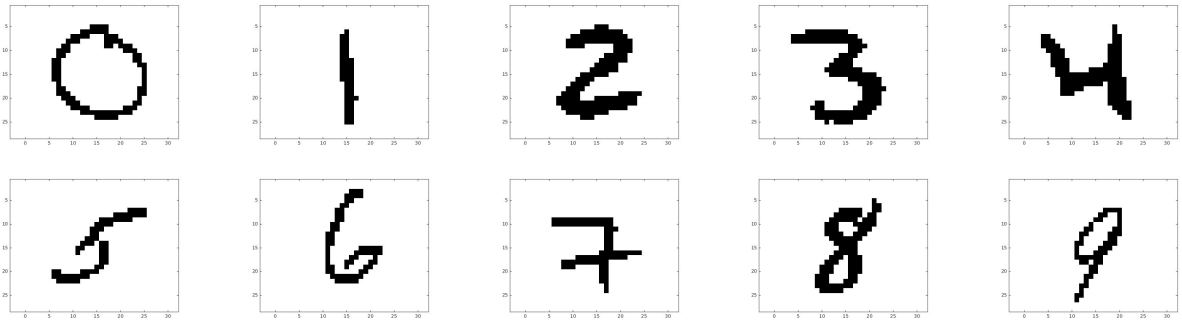

Figure 1: A sample from the binary handwritten digit data. Each pixel of a 16×16 image is either 0 (depicted in white) or 1 (depicted in black).

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 (5). 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 $\mathcal{T}_m$, $m = 1, \ldots, M$.

5.3 Results obtained with Sylla & Girard kernel

We first investigate the use of Sylla & Girard similarity measure (4) when plugged into (5). The CCR are computed for $\alpha \in \{0, 0.1, \ldots, 1\}$ and for several proportions τ in 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 have an effect on the classification. This is different to 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. 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 [38].

5.4 Results obtained with the 76 kernels from [34]

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 [34] with the exponential kernel (5). 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$\in [78.65\%, 89.70\%]$. It is also interesting to note that 8 kernels out of the 76 of [34] 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

Finally, the proposed classification method is compared to the Random Forest method (`RandomForest` package, version 4.6-10 from `R` software) and the SVM method (library `libsvm`, version 3.2 from `Matlab`). The "one-against-all" implementation of the SVM classification method is used. To this aim, we limit ourselves to the use of Sylla & Girard kernel in pgpDA and SVM methods. It appears in Table 3 that on the verbal autopsy dataset, pgpDA method yields better results than SVM and Random Forest on the test set. Since the CCR obtained with Random Forest is larger on the learning set, one can suspect that Random Forest overfits this dataset. One can also observe that the CCR associated with ppgDA slightly depends on α (CCR$\in$ $[72.88\%, 76.43\%]$ whereas CCR associated with SVM is very sensitive to α (CCR$\in$ $[61.19\%, 74.57\%]$. At the opposite, SVM and Random Forest yield better results than pgpDA on the handwritten digits dataset. This may due to the small number of classes ($L = 10$ here, $L = 22$ in the previous situation). The CCR associated with ppgDA is however satisfying, it is larger than 87% whatever the value of α is.

6 Conclusion

This work was motivated by two facts: First, numerous binary similarity measures have been used in various scientific fields. Second, model-based mixture models provide a coherent response to the problem of classification with probabilistic interpretation and natural multi-class support. Basing on these remarks, our main contribution is the proposal of a new classification method combining advantages from both model-based mixture models and binary similarity measures. The method enjoys offer 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 binary classification problems. As a by-product of this work, some new similarity measures are proposed to unify the existing literature.

Appendix: Proofs

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

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

Thus, $(\tilde{M}_k)_{\ell,\ell'} := \tilde{\rho}_k(x_\ell, x_{\ell'})/n_k = \eta(M_k)_{\ell,\ell'}$ for all $(\ell,\ell') \in \{1,\ldots,n_k\}^2$. 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 $\tilde{\beta}_{k1},\ldots,\tilde{\beta}_{kn_k}$. Clearly, $\tilde{\lambda}_{kj} = \eta\lambda_{kj}$ and $\tilde{\beta}_{kj} = \pm\beta_{kj}$ for all $(j,k) \in \{1,\ldots,n_k\}^2$. 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

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

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

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

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

Proof of Lemma 3. 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 (2).

– The second step consists in showing that $1/S_2$ defines a proper kernel classification method. Let us focus on the case where $0 \geq \alpha', \beta' < \theta'$, 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 [35], Proposition 3.22 that $S_{3,N}$ defines a proper kernel classification method for all $N > 0$. As a consequence, $S_{3,N}$ verifies condition (2) for all $N > 0$. Letting $N \to \infty$, one thus has that $1/S_2$ defines a proper kernel classification method.

– Finally, in view of [35], Proposition 3.22 and Proposition 3.24, it follows that K defines a proper kernel classification method. ■

References

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

[2] V. Batagelj & M. Bren. Comparing resemblance measures. *Journal of Classification*, **12**, 73–90, 1995.

[3] F.B. Baulieu. A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 233–246, 1989.

[4] N. Bouguila, D. Ziou & J. Vaillancourt. Novel mixtures based on the Dirichlet distribution: application to data and image classification. *In Machine Learning and Data Mining in Pattern Recognition*, pages 172–181, Springer, 2003.

[5] C. Bouveyron & C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, **22**, 301–324, 2012.

[6] C. Bouveyron, M. Fauvel & S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, to appear, 2015.

[7] C. Bouveyron, S. Girard & C. Schmid, C. High-dimensional discriminant analysis. *Communications in Statistics - Theory and Methods*, **36**, 2607–2623, 2007.

[8] C. Bouveyron, S. Girard & C. Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, **52**, 502–519, 2007.

[9] P. Byass, D.L. Huong & H.V. Minh. A probabilistic approach to interpreting verbal autopsies: Methodology and preliminary validation in Vietnam. *Scand. J. Public Health*, **31**(62):32–37, 2003.

[10] R. Cattell. The scree test for the number of factors. *Multivar. Behav. Res.* **1**(2), 245–276, 1966.

[11] G. Celeux & G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of classification*, **8**, 157–176, 1991.

[12] M.M. Dundar & D.A. Landgrebe. Toward an optimal supervised classifier for the analysis of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **42**(1), 271–277, 2004.

[13] F. Forbes & D. Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tail-weight: application to robust clustering. *Statistics and Computing*, **24**(6), 971–984, 2014.

[14] D. Wraith & F. Forbes. Location and scale mixtures of Gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics and Data Analysis*, **90**, 61–73, 2015.

[15] B.C. Franczak, R.P. Browne & P.D. McNicholas. Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** (6), 1149–1157, 2014.

[16] L.A Goodman & W.H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732–764, 1954.

[17] L.A Goodman & W.H. Kruskal. Measures of association for cross classifications II. Further discussion and references. *Journal of the American Statistical Association*, **54**, 35–75, 1959.

[18] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Anal. Appl.*, **5**(2), 168–179, 2002.

[19] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, **8**, 2551–2594, 2007.

[20] T. Hofmann, B. Schölkopf & A. Smola, A. Kernel methods in machine learning. *Annals of Statistics*, **36**(3), 1171–1220, 2008.

[21] D.L. Huong, H.V. Minh & P. Byass. Applying verbal autopsy to determine cause of death in rural Vietnam. *Scand. J. Public Health*, **31**(62), 19–25, 2003.

[22] Y. LeCun, L. Bottou, Y. Bengio & P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of IEEE*, **86**(11), 2278–2324, 1998.

[23] P. Jaccard. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaudoise Sci. Nat.*, **37**, 547–579, 1901.

[24] S. Lee & G. McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, **24**(2), 181–202, 2013.

[25] T.I. Lin. Robust mixture modeling using multivariate skew t-distribution. *Statistics and Computing*, **20**, 343–356, 2010.

[26] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[27] G. McLachlan, D. Peel & R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–388, 2003.

[28] P. McNicholas & B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296, 2008.

[29] S. Mika, G. Ratsch, J. Weston, B. Schölkopf & K.R. Müllers. Fisher discriminant analysis with kernels. *In: Neural Networks for Signal Processing (NIPS)*, pp. 41–48, 1999.

[30] A. Montanari & C. Viroli. Heteroscedastic factor mixture analysis. *Statistical Modeling*, **10**, 441–460, 2010.

[31] T.B. Murphy, N. Dean & A.E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, **4**, 219–223, 2010.

[32] E. Pekalska & B. Haasdonk. Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(6), 1017–1032, 2009.

[33] B. Scholkopf & A.J. Smola. *Learning with Kernels*, The MIT Press, 1990.

[34] C. Seung-Seok, C. Sung-Hyuk & C. Tappert. A survey of binary similarity and distance measures. *Systemics, Cybernetics and Informatics*, **8**, 43–48, 2010.

[35] J. Shawe-Taylor & N. Cristianini. *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[36] B.C. Reeves & M.A. Quigley. A review of data-derived methods for assigning causes of death from verbal autopsy data. *Int. J. Epidemiology*, **26**, 1080–1089, 1997.

[37] P.H.A. Sneath & R.R. Sokal. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W.H. Freeman and Company, San Francisco, 1973.

[38] S. Sylla, S. Girard, A. Diongue, A. Diallo & C. Sokhna. Classification supervisée par modèle de mélange: Application aux diagnostics par autopsie verbale, *46èmes Journées de Statistique organisées par la Société Française de Statistique*, Rennes, 2014.

[39] A. Tversky. Feature of similarity, *Psychological Review*, **84**, 327–352, 1977.

[40] F. Vilca, N. Balakrishnan & C. Zeller. Multivariate skew-normal generalized hyperbolic distribution and its properties. *Journal of Multivariate Analysis*, **128**, 73–85, 2014.

[41] J. Wang, J. Lee & C. Zhang. Kernel trick embedded Gaussian mixture model. *In: Proceedings of the 14th International Conference on Algorithmic Learning Theory*, pp. 159–174, 2003.

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

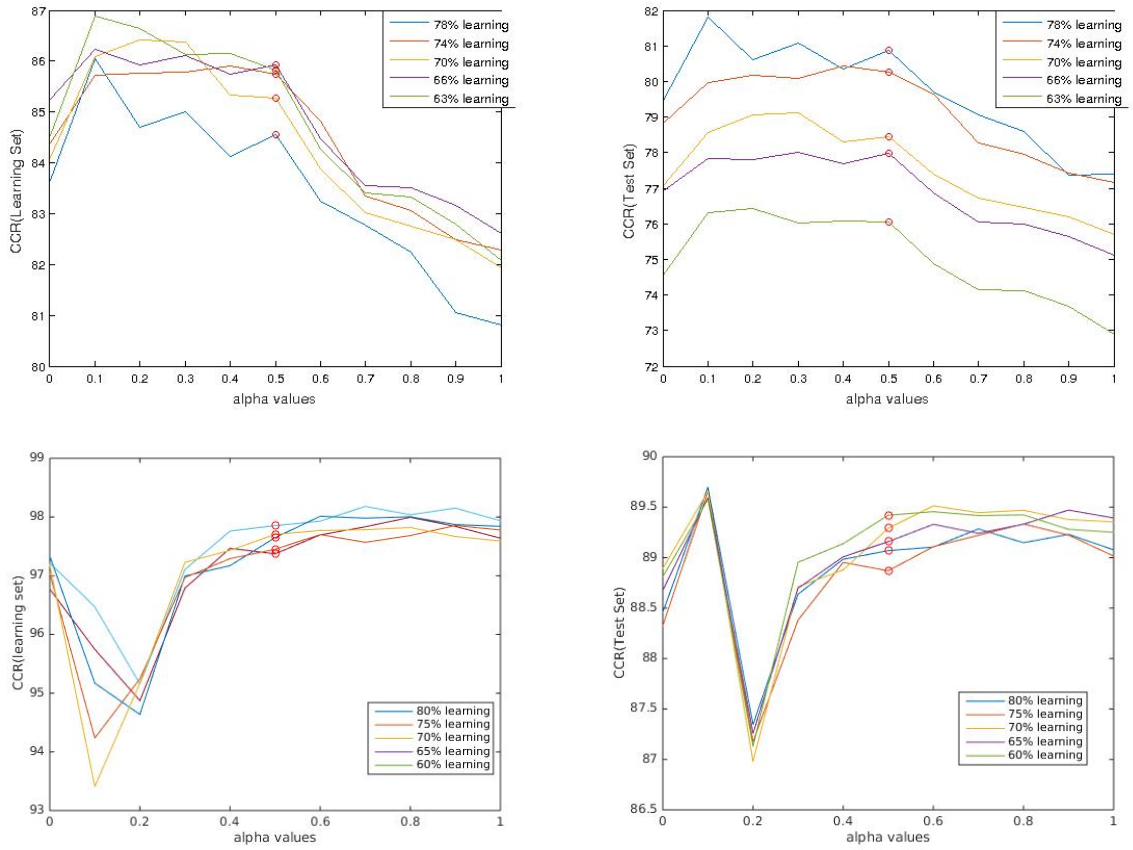

Figure 2: Correct Classification Rate (CCR) obtained with Sylla & Girard kernel (4, 5) for $\alpha \in \{0, 0.1, \ldots, 1\}$ and several proportions τ. The results obtained with the RBF kernel ($\alpha = 0.5$) are emphasized by a red circle. Left: CCR computed on the learning set, Right: CCR computed on the test set. Top: results obtained on the verbal autopsy data, bottom: results obtained on the handwritten digit data.

Kernel	α	σ	`threshold` t	**CCR (learning set)**	**CCR (test set)**	**equation**
Euclid		4	0.60	87.99	83.82	(16)
Pearson		10	0.95	87.72	83.25	(51)
Hellinger		6	0.60	87.68	83.21	(29,30)
Dice		2	0.60	87.32	83.00	(2,3,5)
3w-Jaccard		2	0.75	87.21	82.87	(4)
Ochia1		2	0.60	87.15	82.77	(33,38)
Gower & Legendre		4	0.80	86.61	82.64	(8,11)
Roger & Tanimoto		2	0.65	85.89	82.39	(9)
Sylla & Girard	0.1	1.9	0.90	85.81	81.50	
Sylla & Girard	0.3	2.2	0.85	85.47	81.46	
Sylla & Girard = RBF	0.5	1.4	0.80	85.05	81.35	(15,17,...,23)
Sylla & Girard	0.2	1.8	0.80	85.58	81.11	
Godman & Kruskal		4	0.95	84.28	80.77	(69)
Sylla & Girard	0.4	2.5	0.80	84.67	80.64	
Sokal & Sneath 5		4	0.95	84.72	80.49	(57)
Sylla & Girard	0.6	3.09	0.80	83.19	79.61	
Sylla & Girard	0.7	3.34	0.95	82.96	79.47	
Sokal & Sneath1		2	0.05	83.37	78.65	(6)
Kernel	α	σ	`threshold` t	**CCR (learning set)**	**CCR (test set)**	**equation**
Hellinger		8	0.5	97.57	89.70	(29,30)
Euclid		8	0.5	97.54	89.67	(16)
Sylla & Girard	0.1	3.16	1	92.29	89.58	
Sylla & Girard	0.6	6.19	0.5	97.52	89.46	
Sylla & Girard	0.7	6.69	0.5	97.47	89.41	
Dice		2	0.5	97.40	89.41	(2,3,5)
Ochia1		2	0.5	97.36	89.38	(33,38)
Sylla & Girard = RBF	0.5	5.65	0.5	97.49	89.38	(15,17,. . . ,23)
Roger & Tanimoto		2	0.4	97.29	89.35	(9)
Sylla & Girard	0.8	8	0.5	97.38	89.31	
Sylla & Girard	1	8	8	92.29	89.26	(9)
3w-Jaccard		4	0.5	97.28	89.25	(4)
Sylla & Girard	0.9	7.15	0.5	97.27	89.23	
Jaccard		4	0.4	97.23	89.21	(1)
Gower & Legendre		10	0.8	97.36	89.14	(8,11)
Sylla & Girard	0.4	6.3	0.5	97.21	89.10	
Sylla & Girard	0.3	5.4	0.5	96.86	88.67	
Sylla & Girard	0.2	4.4	0.45	97.94	86.82	

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 [34].

	pgpDA		**SVM**		**Random Forest**	
α	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**
0.1	86.89	76.30	85.32	74.57	-	-
0.2	86.64	76.43	79.87	70.77	-	-
0.3	86.13	76.01	79.49	70.42	-	-
0.4	86.15	76.07	76.03	67.94	-	-
0.5	85.82	76.05	72.66	65.35	-	-
0.6	84.26	74.87	70.28	63.52	-	-
0.7	83.40	74.15	69.16	62.55	-	-
0.8	83.32	74.11	68.66	62.21	-	-
0.9	82.80	73.69	68.20	61.72	-	-
1	82.08	72.88	67.63	61.19	-	-
-	-	-	-	-	87.25	67.68

	pgpDA		**SVM**		**Random Forest**	
α	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**
0.1	93.40	89.57	100.00	93.13	-	-
0.2	95.17	87.13	99.99	97.55	-	-
0.3	97.22	88.95	99.96	97.85	-	-
0.4	97.43	89.13	99.75	97.69	-	-
0.5	97.70	89.42	99.44	97.37	-	-
0.6	97.77	89.45	99.26	97.20	-	-
0.7	97.78	89.41	99.10	97.02	-	-
0.8	97.82	89.42	98.29	96.25	-	-
0.9	97.66	89.28	98.05	96.00	-	-
1	97.59	89.25	97.67	95.66	-	-
-	-	-	-	-	100	93.92

Table 3: Correct Classification Rate (CCR) on the verbal autopsy dataset (top) and on the handwritten digit dataset (bottom). The Random Forest method is compared to the Sylla & Girard kernel plugged into pgpDA and SVM classification methods. The train set includes $\tau = 63\%$ individuals from the initial dataset.