CLASSIFICATION USING HIGH ORDER DISSIMILARITIES IN NON-EUCLIDEAN SPACES

Helena Aidos[1], Ana Fred[1] and Robert P. W. Duin[2]
[1]*Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal*
[2]*Faculty of Electrical Engineering, Mathematics and Computer Sciences, Delft University of Technology, Delft, The Netherlands*

Keywords: Dissimilarity increments, Maximum a posteriori, Classification, Gaussian mixture, Non-Euclidean space.

Abstract: This paper introduces a novel classification algorithm named MAP-DID. This algorithm combines a maximum a posteriori (MAP) approach using the well-known Gaussian Mixture Model (GMM) method with a term that forces the various Gaussian components within each class to have a common structure. That structure is based on higher-order statistics of the data, through the use of the dissimilarity increments distribution (DID), which contains information regarding the triplets of neighbor points in the data, as opposed to typical pairwise measures, such as the Euclidean distance. We study the performance of MAP-DID on several synthetic and real datasets and on various non-Euclidean spaces. The results show that MAP-DID outperforms other classifiers and is therefore appropriate for classification of data on such spaces.

1 INTRODUCTION

Classification deals with algorithmic methodologies for assigning a new input data to one of the known classes. There are numerous classifiers with different strategies, like *k* nearest neighbor, neural networks, support vector machines, Parzen windows (Duda et al., 2001; Hastie et al., 2009).

This paper introduces a new maximum a posteriori (MAP) classifier based on the Gaussian Mixture Model (GMM). This novel classifier (MAP-DID) introduces an extra factor on the likelihood containing information about higher-order statistics of the data, through the use of the distribution of their dissimilarity increments (Aidos and Fred, 2011).

2 DISSIMILARITY REPRESENTATIONS

Sometimes it is useful to describe the objects using a dissimilarity representation, a square matrix containing the dissimilarities between all pairs of objects. To use the typical classifiers, we need to build a vector space based on the relations given by the dissimilarity matrix. In (Duin and Pekalska, 2008), two strategies are considered to obtain vector spaces: pseudo-Euclidean spaces and dissimilarity spaces.

2.1 Pseudo-Euclidean Spaces (PES)

The PES is given by the Cartesian product of two real spaces: $\mathcal{E} = \mathbb{R}^p \times \mathbb{R}^q$. A vector $\mathbf{x} \in \mathcal{E}$ is represented as an ordered pair of two real vectors: $\mathbf{x} = (\mathbf{x}^+, \mathbf{x}^-)$. This space is equipped with a pseudo-inner product, such that $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{E}} = \mathbf{x}^T \mathcal{J}_{pq} \mathbf{y}$, where $\mathcal{J}_{pq} = [I_{p\times p} \ 0; 0 \ -I_{q\times q}]$. Alternatively, if x_i^+ and x_i^- represent the components of $\mathbf{x}^+$ and $\mathbf{x}^-$, then $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{E}} = \sum_{i=1}^{p} x_i^+ y_i^+ - \sum_{i=1}^{q} x_i^- y_i^-$.

Although this pseudo-inner product is symmetric and linear in its first argument, it is not positive definite. Thus, if one constructs the Gram matrix, $\mathbf{G}$, from the data patterns $\mathbf{x}_i$ as $\mathbf{G}_{ij} = \mathbf{x}_i^T \mathbf{x}_j$, then $\mathbf{G}$ may not be positive semidefinite in the PES (Pekalska, 2005). $\mathbf{G}$ is symmetric in the PES, so it has an eigendecomposition of $\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^T$; but, its eigenvalues can be negative. Note that a new dataset can be built up from $\mathbf{G}$ through $\mathbf{X} = \mathbf{V}|\mathbf{D}|^{1/2}$, where matrix $\mathbf{X}$ contains the vector representations of the new patterns in the PES.

In (Duin et al., 2008; Duin and Pekalska, 2008), several variants of PES are considered. In this paper, we also consider the following spaces.

- ***Pseudo-Euclidean Space (PES):*** This is a $(p+q)$-dimensional PES defined by $p+q$ eigenvectors. One keeps the p largest positive eigenvalues

Aidos H., Fred A. and P. W. Duin R. (2012).
CLASSIFICATION USING HIGH ORDER DISSIMILARITIES IN NON-EUCLIDEAN SPACES.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 306-309
DOI: 10.5220/0003779503060309

and the q negative eigenvalues that have the highest absolute value. Each direction is scaled by the magnitude of the corresponding eigenvalue.

- ***Positive Pseudo-Euclidean Space (PPES):*** This p-dimensional space is defined as PES, but only the p largest positive eigenvalues are kept.
- ***Negative Pseudo-Euclidean Space (NPES):*** This q-dimensional space is defined as PES, but only the q largest negative eigenvalues (in magnitude) are kept; no positive eigenvalues are used.
- ***Corrected Euclidean Space (CES):*** In CES, a constant is added to all the eigenvalues (positive and negative) to ensure that they all become positive. This constant is given by $2|a|$, where a is the negative eigenvalue with the largest absolute value.

2.2 Dissimilarity Spaces (DS)

We consider four more spaces constructed in the following way: we compute the pairwise Euclidean distances between data points of one of the spaces defined above. These distances are new feature representations of $\mathbf{x}_i$. Note that the dimension of the feature space is equal to the number of points.

Since our classifier suffers from the curse of dimensionality, we must reduce the number of features; there are several techniques for that (Hastie et al., 2009). We chose k-means to find a number of prototypes $k < N$. k is selected as a certain percentage of $N/2$, and the algorithm is initialized in a deterministic way as described in (Su and Dy, 2007). After the k prototypes are found, the distances from each point $\mathbf{x}_i$ to each of these prototypes are used as their new feature representations. This defines four new spaces, which are named as *Dissimilarity Pseudo-Euclidean Space (DPES)*, *Dissimilarity Positive Pseudo-Euclidean Space (DPPES)*, *Dissimilarity Negative Pseudo-Euclidean Space (DNPES)* and *Dissimilarity Corrected Euclidean Space (DCES)*.

3 THE MAP-DID ALGORITHM

In this section, dissimilarities between patterns in the eight previously defined spaces are computed as Euclidean distances.

3.1 Dissimilarity Increments Distribution (DID)

Let X be a set of patterns, and $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ a triplet of nearest neighbors belonging to X, where $\mathbf{x}_j$ is the nearest neighbor of $\mathbf{x}_i$ and $\mathbf{x}_k$ is the nearest neighbor of $\mathbf{x}_j$, different from $\mathbf{x}_i$. The *dissimilarity increment* (DI) (Fred and Leitão, 2003) between these patterns is defined as $d_{inc}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \left|d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k)\right|$. This measure contains information different from a distance: the latter is a pairwise measure, while the former is a measure for a triplet of points, thus a measure of higher-order dissimilarity of the data.

In (Aidos and Fred, 2011) the DIs distribution (DID) was derived under the hypothesis of Gaussian distribution of the data and it was written as a function of the mean value of the DIs, λ. Therefore, the DID of a class is given by

$$p_{d_{inc}}(w;\lambda) = \frac{\pi\beta^2}{4\lambda^2} w \exp\left(-\frac{\pi\beta^2}{4\lambda^2}w^2\right) + \frac{\pi^2\beta^3}{8\sqrt{2}\lambda^3} \times \left(\frac{4\lambda^2}{\pi\beta^2} - w^2\right) \exp\left(-\frac{\pi\beta^2}{8\lambda^2}w^2\right) \operatorname{erfc}\left(\frac{\sqrt{\pi}\beta}{2\sqrt{2}\lambda}w\right), \quad (1)$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function, and $\beta = 2 - \sqrt{2}$.

3.2 MAP-DID

Consider that $\{\mathbf{x}_i, c_i, inc_i\}_{i=1}^N$ is our dataset, where $\mathbf{x}_i$ is a feature vector in $\mathbb{R}^d$, c_i is the class label and inc_i is the set of increments yielded by all the triplets of points containing $\mathbf{x}_i$. We assume that a class c_i has a single statistical model for the increments, with an associated parameter λ_i. This DID, described above, can be seen as high-order statistics of the data since it has information of a third order dissimilarity of data.

For example, we generate a 2-dimensional Gaussian with 1000 points; it has zero mean and covariance the identity matrix (figure 1 left). We also generate a 2-dimensional dataset with 1000 points, where 996 points are in the center and there are four off-center points at coordinates $(\pm a, 0)$ and $(0, \pm a)$, where a is such that the covariance is also the identity matrix (figure 1 right). We compute the DIs for each dataset and look at their histograms (figure 1).

Although the datasets have the same mean and covariance matrix, the two DIs distributions are very different from each other. Therefore, the DIs can be seen as a measure of higher-order statistics: the two distributions under consideration have exactly the same mean and variance, but their DIDs are vastly different.

So, we design a maximum a posteriori (MAP) classifier that combines the Gaussian Mixture Model (GMM) and the information given by the increments, assuming that $\mathbf{x}_i$ and inc_i are conditionally independent given c_j. We used a prior given by $p(c_j) = |c_j|/N$, with $|c_j|$ the number of points of class j, and the likelihood $p(\mathbf{x}_i, inc_i|c_j) = p(\mathbf{x}_i|c_j)p(inc_i|c_j)$.

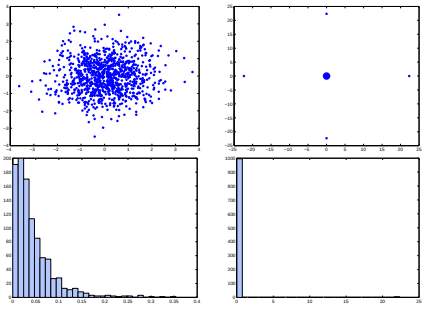

Figure 1: Two simple datasets with zero mean and a covariance given by the identity matrix, but with vastly different DIs. *Left:* Gaussian data. *Right:* dataset with 996 points at the origin and four off-center points. Corresponding histograms of the DIs. Note that in the right histogram there are four non-zero increments and 996 zero increments.

The class-conditional density of the vector $\mathbf{x}_i$ follows a GMM given by $p(\mathbf{x}_i|c_j) = \sum_{l=1}^{K} \alpha_l p(\mathbf{x}_i|\Sigma_l, \mu_l)$, with K the number of Gaussian components determined for class c_j, α_l the weight of each Gaussian component and $p(\mathbf{x}_i|\Sigma_l, \mu_l)$ the Gaussian distribution. We obtained the parameters α_l, Σ_l and μ_l using the GMM described in (Figueiredo and Jain, 2002).

The class-conditional density of the set of increments where $\mathbf{x}_i$ belongs is given by $p(inc_i|c_j) = \frac{1}{M}\sum_{n=1}^{M} p(inc_i^n|c_j)$, where M is the number of increments of the set inc_i, inc_i^n is the n-th increment of that set, and $p(inc_i|c_j) = p(inc_i|\lambda_j)$ is the DID given in equation (1). We thus consider a statistical model for increments with parameter λ_j for each class.

4 EXPERIMENTAL RESULTS AND DISCUSSION

In this section we compare MAP-DID to other classifiers (1-nearest neighbor (1-NN), nearest-mean (NM), Parzen window and a linear support vector machine (SVM)). We use 13 datasets, of which 2 are synthetic and 11 are real-world data[1].

For each of the classifiers, we use a 10-fold cross-validation scheme to estimate classifier performance. Figures 2 and 3 present the results for the average classification error. The values of p and q eigenvectors, and k prototypes, used to construct the spaces described in Section 2, are in Table 1.

The MAP-DID is the algorithm with the lowest error rate. This is true for the vast majority of all the possible dataset-space pairs. Thus, if any of these spaces are to be used for classification, the MAP-DID is a good choice for classification algorithm.

Some other interesting points should be emphasized: for the real-world datasets it is interesting to note that the results are not very different between the PES, PPES and CES spaces, all of which take into account the positive portion of the space. Conversely, the NPES results are considerably worse than those

[1]See http://prtools.org/disdatasets/ for a full description of the datasets and the MATLAB toolboxes containing the classifiers used for comparison.

Table 1: Number of eigenvectors and prototypes used to construct the spaces described in Section 2.

Dataset	p	q	k
Balls3d	3	7	10
Balls50d	18	5	18
CatCortex	2	2	2
CoilDelftSame	8	4	7
CoilYork	8	5	7
DelftGestures	11	2	13
Protein	6	2	5
Zongker	14	3	20
Chickenpieces	8	3	9

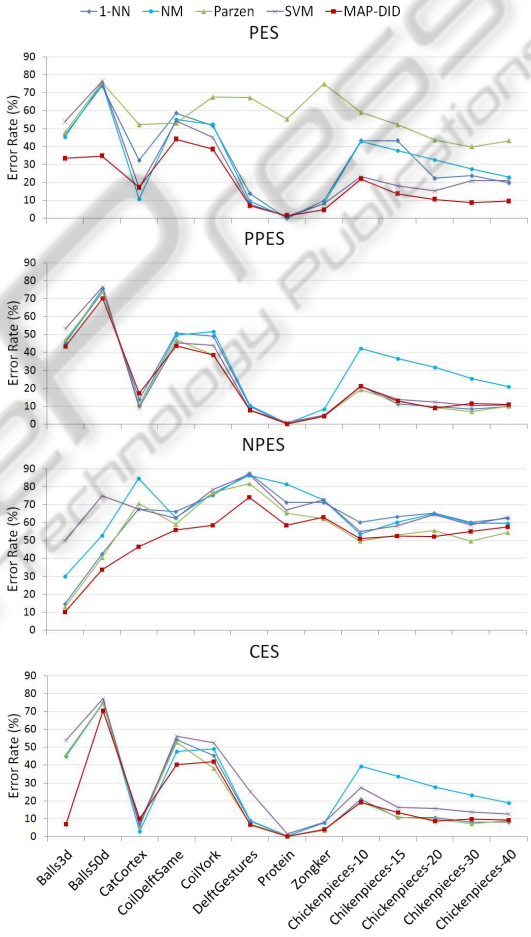

Figure 2: Classification error rate on the four pseudo-Euclidean spaces considered in Section 2.1.

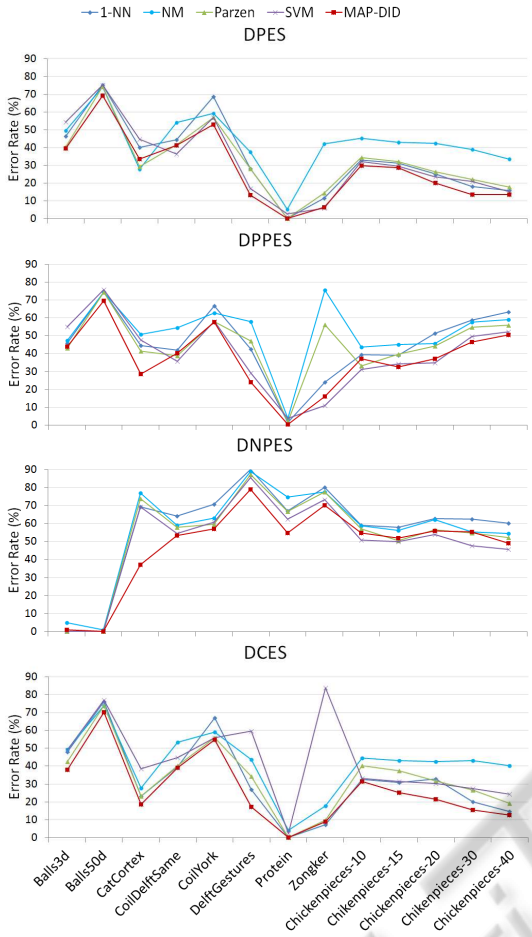

Figure 3: Classification error rate on the four dissimilarity spaces considered in Section 2.2.

three, which indicates that this negative space contains little information for classification purposes.

Another interesting point is that in the dissimilarity spaces (figure 3), neither the positive (DPPES) nor the negative (DNPES) spaces contain all the information; instead, the union of the information contained in those two spaces (DPES or DCES) yields much better results than either of them separately.

It was necessary to reduce the dimensionality of the data to generate the dissimilarity spaces (figure 3). This reduction was accomplished through *k*-means, by computing the distances from the data patterns to the estimated prototypes. However, many other techniques could be used for dimensionality reduction, and it is possible that some of those techniques would yield an improvement on the results for these spaces.

One aspect not considered here is the metricity and euclideaness of datasets (Duin and Pekalska, 2008). These properties may help us identify the situations where MAP-DID performs well.

5 CONCLUSIONS

We have presented a novel maximum a posteriori (MAP) classifier which uses the dissimilarity increments distribution (DID). This classifier, called MAP-DID, can be interpreted as a Gaussian Mixture Model with an operator that forces a class to have a common increment structure, even though each Gaussian component within a class can have distinct means and covariances. Experimental results have shown that MAP-DID outperforms multiple other classifiers on various datasets and feature spaces.

In this paper we focused on Euclidean spaces derived from non-Euclidean data. This might suggest that MAP-DID could perform well when applied to originally Euclidean data. This is a topic which will receive more investigation in the future.

ACKNOWLEDGEMENTS

This work was supported by the FET programme within the EU FP7, under the SIMBAD project contract 213250; and partially by the Portuguese Foundation for Science and Technology, scholarship number SFRH/BD/39642/2007, and grant PTDC/EIA-CCO/103230/2008.

REFERENCES

Aidos, H. and Fred, A. (2011). On the distribution of dissimilarity increments. In *IbPRIA*, pages 192–199.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons Inc., 2nd edition.

Duin, R., Pekalska, E., Harol, A., Lee, W.-J., and Bunke, H. (2008). On euclidean corrections for non-euclidean dissimilarities. In *SSPR/SPR*, pages 551–561.

Duin, R. P. and Pekalska, E. (2008). On refining dissimilarity matrices for an improved nn learning. In *ICPR*.

Figueiredo, M. and Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Fred, A. and Leitão, J. (2003). A new cluster isolation criterion based on dissimilarity increments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):944–958.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Pekalska, E. (2005). *Dissimilarity Representations in Pattern Recognition: Concepts, Theory and Applications*. PhD thesis, Delft University of Technology, Delft, Netherland.

Su, T. and Dy, J. G. (2007). In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intelligent Data Analysis*, 11(4):319–338.