A New Method for Prototype Selection in Dissimilarity Spaces

Yenisel Plasencia Calaña, Edel García Reyes, Robert P. W. Duin, Mauricio Orozco-Alzate

Advanced Technologies Application Centre, 7a # 21812, Siboney, Playa, Havana - 12200, Cuba.

{yplasencia,egarcia}@cenatav.co.cu, R.P.W.Duin@tudelft.nl, morozcoa@unal.edu.co

Temática: reconocimiento de patrones basado en disimilitud

Delegación de base: La Habana

Abstract.
The dissimilarity representation is a powerful tool for representing objects like images and graphs where the extraction of good features can be difficult, expensive, or impossible. For computing a dissimilarity representation some ingredients are needed, a dissimilarity measure for the problem at hand and a set of prototype objects, that can be selected from the training set. Then, for each object, dissimilarities are measured with the set of prototypes and the representation is the vector of obtained dissimilarities. The length of these vectors is equal to the size of the set of prototypes. A small representation set is desirable for many real world applications since it will lead to lower computational costs in the classification stage. In this paper we present a new method for selecting a small set of prototypes capable of maintaining the cost of classification in dissimilarity spaces as low as possible without a major loss in accuracy. Experimental results on four datasets show the validity of the proposed approach.

Resumen.
La representación por disimilitud es una herramienta potente para representar objetos como imágenes y grafos donde la extracción de buenos rasgos puede ser difícil, costosa o imposible. Para obtener una representación por disimilitud algunos ingredientes son necesarios, una medida de disimilitud para el problema en cuestión y un conjunto de objetos prototipos, que pueden seleccionados del conjunto de entrenamiento. Luego, para cada objeto, las disimilitudes son medidas con el conjunto de prototipos y la representación es el vector de disimilitudes obtenidas. La longitud de estos vectores es igual al tamaño del conjunto de prototipos. Un conjunto de representación pequeño es deseable para muchas aplicaciones del mundo real, ya que esto conllevará a menores costos computacionales en la etapa de clasificación. En este artículo presentamos un nuevo método para seleccionar un conjunto de prototipos pequeño capaz de mantener el costo de clasificación en espacios de disimilitud lo más bajo posible sin una perdida mayor de la precisión. Los resultados experimentales en cuatro bases de datos muestran la validez del enfoque propuesto.

1. Introduction

The representation of objects in Euclidean or metric spaces with the use of statistical classifiers in these spaces has been the most widely used strategy to approach pattern recognition. This strategy has demonstrated to be useful for a variety of problems, and it has many advantages such as the great amount of statistical tools available for Euclidean vector spaces. Other trends for representation have arisen as a very promising alternative such as structural representations and dissimilarity representations. Examples of structural representation [1-3] are graphs and strings, which are robust and flexible approaches. One drawback of this representation is that there is a lack of tools for classification, opposite to statistical pattern recognition. So, bridging the gap between the two approaches has been an aspect of research, in order to make the tools of statistical pattern recognition suitable for the classification of structural representations such as graphs. One way to do this is to use dissimilarity representations on top of the structural representation[2]. Then, the classification can be carried out in dissimilarity spaces.

Dissimilarity representations were proposed by Pekalska et. al. [4]. They can be computed directly on raw data, but also on top of feature and structural representations. Then, the classification of objects can be done by different approaches: classification by the *k* nearest neighbor (*k*-NN) rule, classification in dissimilarity spaces, classification in Pseudo-Euclidean spaces, and, if the dissimilarity is turned into a similarity by algebraic operations, the classification can be carried out by support vector machines. In this paper we restrict ourselves to the classification of objects in the dissimilarity space.

The dissimilarity space is generated by a set of prototypes using a dissimilarity measure for the problem at hand. The set of prototypes can be selected from the training set or can be a totally different set. Then, for each training object, dissimilarities are measured with the set of prototypes and the representation is the vector of obtained dissimilarities. The length of these vectors is equal to the size of the set of prototypes. Statistical classifiers can be trained on the training objects in the dissimilarity space. A small representation set is desirable for many real world applications since it will lead to lower computational costs in the classification stage, this can be achieved by means of prototype selection. In this paper we propose a new method for finding a small representation set that allows one to obtain a good compromise between computational costs and classification accuracy in the dissimilarity space. For prototype selection in dissimilarity spaces several methods have been proposed[5, 6]. Some of them requires an initial feature vector space available [6], but others work directly with dissimilarity data [5].

2. Related concepts and methods

2.1 Dissimilarity space

The dissimilarity space was proposed by Pekalska et. al. [4]. It was postulated as a Euclidean vector space implying the possibility of using several statistical classifiers there. A representation set $R = \{r_1, r_2, \dots, r_k\}$ is defined. The objects belonging to this set are called prototypes and can be chosen based on some criterion or even at random. Let T be the training set, R may or may not overlap with T. Once we have R, the dissimilarities of the objects in T to the objects in R are computed. Any object x is represented by a vector of dissimilarities d_x to the objects in R:

$$x = [d(x,r_1)d(x,r_2)\dots d(x,r_k)]. \qquad (1)$$

The cardinality of R determines the dimension of the space, and each coordinate value of a point corresponds to the dissimilarity with some prototype. The number of prototypes allows one to control the computational cost and to find a tradeoff between classification accuracy and computational efficiency.

2.2 Prototype selection

For dissimilarity representations, the adaptation of prototype selection techniques available for the vector space representation or feature-based approach has been investigated showing good results [6]. In the k-NN literature two basic types of algorithms can be identified: prototype generation and prototype selection[4, 5]. Prototype selection searches between existent objects; prototype generation demands the creation of artificial objects in an intermediate vectorial space. We assume in this work that dissimilarities are computed from a matching process and not from an available vectorial representation; therefore, we do not have an intermediate vectorial space where artificial objects can be generated, we only have dissimilarities between the objects.

In [6] the authors compare prototype selection methods for classification in dissimilarity spaces created from feature spaces, showing good results when used with linear and quadratic classifiers. In [5], various techniques were compared such as Kcentres, mode seeking, forward selection (FS) minimizing the 1-NN error on the training set, linear programming, editing and condensing, and a mixture of Kcentres with linear programming. These techniques showed good performances. Other prototype selection methods have been proposed in the graph and string domain [1, 3] such as the center prototype selector method. The methods tackle the question of how to select a good and small representation set for constructing the dissimilarity space.

3. Proposed method

In general, prototype selection methods in dissimilarity spaces can be roughly classified as supervised or unsupervised. The methods can use any of the variety of search strategies available for optimizing a criterion that can be supervised or unsupervised. When the method is supervised (wrapper), it takes into account class labels and usually tries to minimize some classification error. These methods are usually more computationally expensive but they can provide good sets of prototypes. Examples of these methods are the FS minimizing the 1-NN error, and editing and condensing. Unsupervised methods, on the other hand, do not need class labels in the optimization process; they have less computational demands, and are able to provide also good representation sets. Examples of unsupervised methods are Kcentres, mode seeking and center prototype selector. It is still an open issue what is the best method, if it exists, for finding a good representation set.

In this paper, we introduce a new unsupervised method for selecting a representation set for dissimilarity space classification, which consists in a FS minimizing the representation error ($FS + rep\ error$), i.e. the expected distance of objects to the representation set. The criterion is motivated by the assumption that a representation set is suitable if the objects in the training set have similar objects in this representation set. The representation set should be selected in a way that the sum of the dissimilarities of each training object to its closest prototype is minimal:

$$min: \Sigma_{r \in R}\, min_{x \in T}\; d(x,r). \qquad (2)$$

This criterion yields the set of objects that provides the best description of the training set in terms of the given dissimilarities. For that reason we call this criterion the representation error.

4. Experimental results

In this section we investigate the performance on four datasets of the proposed method, $FS + rep\ error$, for prototype selection for dissimilarity space classification. A comparison is made with other prototype selection methods and with the 1-NN classifier on the original dissimilarities and not in the dissimilarity space.

4.1 Datasets

Different dissimilarity datasets were used for the experiments. All of them are multiclass problems. A comparison of different properties of all datasets can be found in Table 1.

Table 1. Properties of the datasets used in the experiments.

Datasets	# classes	# objects per class	Symmetric disimilarity	Metric dissimilarity
ChickenPieces-20-60	5	117,76,96,61,96	no	No
CoilYork	4	4x72	No	No
CoilDelftDiff	4	4x72	Yes	No
Swiss Political debates	10	134,242,184,54,9, 274,109,398,46,11	yes	yes

The dissimilarity dataset Chickenpieces-20-60 is computed from the Chickenpieces image dataset [7]. The images are in binary format representing silhouettes from five different parts of the chicken: wing (117 samples), back (76), drumstick (96), thigh and back (61), and breast (96). From these images the edges are extracted and approximated by segments of length 20 pixels, and a string representation of the angles between the segments is derived. The dissimilarity matrix is composed by edit distances between these strings. The cost function between the angles is defined as the difference in case of substitution and as 60 in case of insertion or deletion.

The CoilYork dataset is composed by dissimilarities between a set of graphs derived from four objects of the COIL database, the graphs are the Delaunay triangulations derived from corner points of the images [8]. The dissimilarity matrix is constructed by graph matching, using the algorithm proposed in [9].

The CoilDelftDiff dissimilarity dataset is also computed from a set of graphs derived from four objects of the COIL database. The graphs are the Delaunay triangulations derived from corner points of the images. Graphs are compared in the eigenspace with a dimensionality determined by the smallest graph in every pairwise comparison by the JoEig approach [8].

The Swiss Political Debates dataset contains a single television discussion from a full collection of more than 70 TV recordings from Switzerland. The classification of scenes and speakers in the video is very important for the automatic analysis of these videos. A first step towards the analysis is made by clustering 400 frames of the video stream in an unsupervised manner. First, video frames are extracted at a rate of 1 Hz. Then, from the raw images of resolution 720×576, color histograms are computed using 64 bins per color band. The color histograms are then concatenated (thus creating 192 feature histogram), and between the extended histograms a Chi square distance is computed with the following equation: $\chi^2(S,M) = \sum_{i=1}^{n} \frac{(S_i - M_i)^2}{(S_i + M_i)}$, where S_i and M_i are the bins corresponding to the $i-th$ position in each of the two image histograms respectively. Class labels were assigned to each cluster and the other video frames were classified with the 1-NN rule taking these classes as reference. The resulting classes were inspected and those classes describing the same scene (e.g. moderator and second guest present in the scene) were joined. Also, true labels were manually assigned to those scenes that were erroneously classified in the beginning. From the videos, we derived a dissimilarity dataset using the Chi square distances. Examples of intra-class variabilities can be seen in Fig. 1.

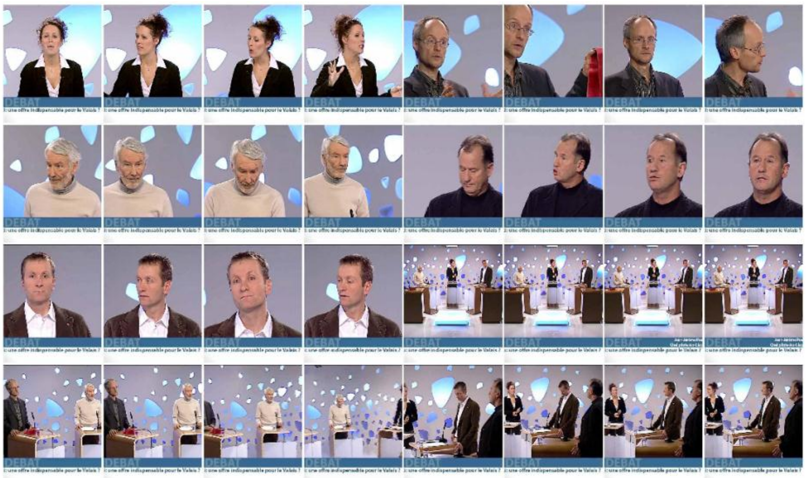

Fig. 1. Example of classes (4 images per class) in the Swiss Political Debates video.

4.2 Results and discussion

As a baseline for the experiments, it is computed the 1-NN classifier in the original dissimilarities. In the dissimilarity space the Linear Discriminant (LD) classifier is used for evaluating the classification results after the prototypes were selected with the different prototype selection methods, we use this classifier instead of other complicated ones such as support vector machines in order to maintain the computational costs of classification as low as possible. The prototype selection method $FS + rep\ error$ is compared to other methods such as the random selection, unsupervised systematic procedures such as Kcentres and center prototype selector; and a supervised FS minimizing the 1-NN error on the training set as in [5]. All the datasets that were not symmetric in their original version, were symmetrized by averaging the dissimilarities in the two directions $d(a,b)$ and $d(b,a)$ in order to make the methods comparable. The datasets were randomly split twenty times into training and test sets taking approximately 50% of the objects in each set. Then, prototypes were selected from the training set, searching from one to twenty prototypes. Classification results were computed for the objects in the test set. Figs. 2-5 show the averaged results over the twenty runs of the experiments on the datasets.

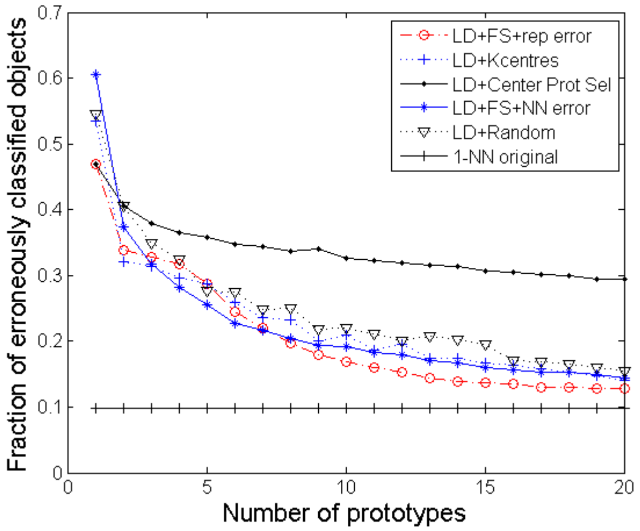

Fig. 2. Classification results in the Chickenpieces20-60 dataset.

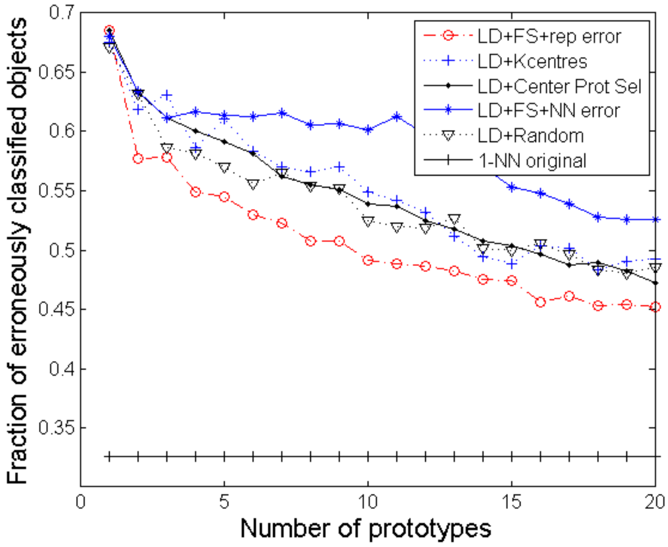

Fig. 3. Classification results in the CoilYork dataset.

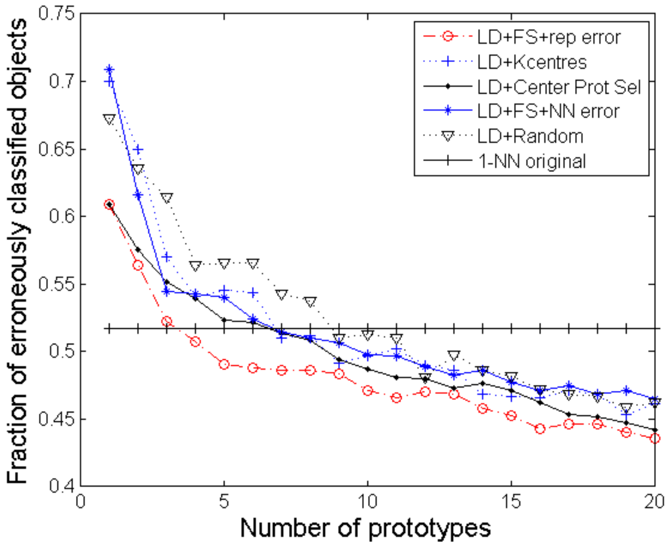

Fig. 4. Classification results in the CoilDelftDiff dataset.

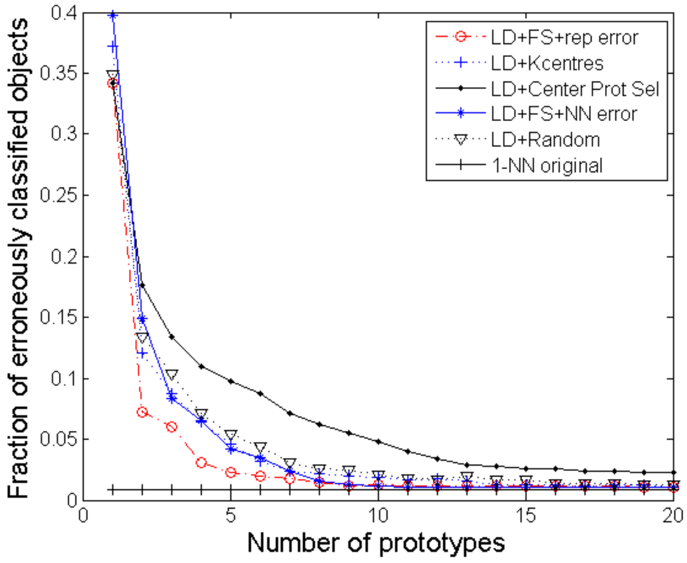

Fig. 5. Classification results in the Swiss Political Debates dataset.

The results shown in Figs. 2-5 are the average errors of the LD in the dissimilarity space for the different prototype selection methods and 1-NN in the dissimilarity matrix over the twenty repetitions. The curves with circles are the ones of the proposed approach. It is important to remark that the 1-NN results are better in most cases but at the cost of measuring the dissimilarities with respect to all the training objects what is computationally demanding for objects like graphs as in CoilYork and CoilDelftDiff datasets. We overcome this by using prototype selection and a compromise between efficiency and accuracy can be achieved. It can be seen from Fig. 2 that in all the experiments the $FS + rep\ error$ leads to equal or better classification results than the random selection and the rest of the unsupervised methods, i.e. Kcentres and center prototype selector method. This can be attributed to the fact that the representation error can be more robust to outliers than the Kcentres because the second is sensitive to outliers in the step of minimization of the maximum distances within clusters. Also, the $FS + rep\ error$ can represent the data distribution better than the Kcentres and the center prototype selector methods.

In the Political Debates dataset with the $FS + rep\ error$ it is possible to obtain only 0.02 of error with only 5 prototypes. This is a very good result for automatic video analysis since for classifying a new image this implies that it is sufficient to measure dissimilarities between the histogram of the new image and the histograms of 5 prototype images leading to an important saving in computational demands. In contrast, the 1-NN needs to measure 732 dissimilarities (the size of the training set) in order to obtain similar results.

Regarding the comparison with the supervised method, it is possible to see from the results that in three (CoilDelftDiff, CoilYork, and Swiss political debates) of the four datasets, the $FS + rep\ error$ equals or outperforms the supervised method. Still, it can be stated that the proposed approach obtains very good results, taking into account that it does not incorporate the previous knowledge of the class labels for training as in the case of the supervised method.

5. Conclusions

In this paper we present a new method for prototype selection in dissimilarity spaces. The method is based on a FS of prototypes minimizing the representation error. Experiments on different dissimilarity data sets showed that the method outperforms other unsupervised methods in the majority of the datasets and reaches similar or better results than a supervised one. This last remark is interesting since usually supervised methods perform better as they benefit from the label information. Another interesting result is the reduction of dimensionality achieved by the proposed method in a video application, where using a very small number of prototypes (i.e. five) it is possible to obtain almost the same results in accuracy as with classification by the 1-NN using all the training objects.

References

1. Barbara Spillmann, M.N., Horst Bunke, Elzbieta Pekalska, Robert P. W. Duin. *Transforming Strings to Vector Spaces Using Prototype Selection*. in *SSPR/SPR*. 2006: Springer.
2. H. Bunke, K.R. *Graph classication based on dissimilarity space embedding*. in *SSSPR*. 2008.
3. Kaspar Riesen, M.N., Horst Bunke. *Graph Embedding in Vector Spaces by Means of Prototype Selection*. in *GbRPR*. 2007: Springer.
4. Elzbieta Pekalska, R.P.W.D., *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. Machine Perception and Artificial Intelligence. 2005: World Scientific Publishing Co., Inc.
5. Elzbieta Pekalska, R.P.W.D., Pavel Paclík, *Prototype selection for dissimilarity-based classifiers*. Pattern Recogn., 2006. **39**(2): p. 189-208.

6. M. Lozano, J.M.S., J. S. Sánchez, F. Pla, E. Pekalska, R. P. W. Duin, *Experimental study on prototype optimisation algorithms for prototype-based classification in vector spaces*. Pattern Recogn., 2006. **39**(10): p. 1827--1838.
7. H. Bunke, U.B., *Applications of approximate string matching to 2D shape recognition.* Pattern Recognition, 1993. **26**(12): p. 1797-1812.
8. Wan-Jui Lee, R.P.W.D. *An Inexact Graph Comparison Approach in Joint Eigenspace*. in *SSPR \& SPR '08: Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*. 2008. Orlando, Florida: Springer-Verlag.
9. Bai Xiao, E.R.H. *Geometric Characterisation of Graphs*. in *ICIAP*. 2005.