A Compact Representation of Multiscale Dissimilarity Data by Prototype Selection

Yenisel Plasencia-Calaña[1,2](✉), Yan Li[2], Robert P.W. Duin[2],
Mauricio Orozco-Alzate[3], Marco Loog[2], and Edel García-Reyes[1]

[1] Advanced Technologies Application Center, 7ma A ♯ 21406, Playa, Havana, Cuba
{yplasencia,egarcia}@cenatav.co.cu

[2] Faculty of Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, Delft, The Netherlands
yanli.grass@gmail.com, r.duin@ieee.org, m.loog@tudelft.nl

[3] Departamento de Informática y Computación,
Universidad Nacional de Colombia - Sede Manizales,
Kilómetro 7 vía al Aeropuerto, Campus La Nubia – Bloque Q, Piso 2, Manizales,
Colombia
morozcoa@unal.edu.co

Abstract. Multiscale information provides an opportunity to improve the outcomes of data analysis processes. However, if the multiscale information is not properly summarized in a compact representation, this may lead to problems related to high dimensional data. In addition, in some situations, it is convenient to define dissimilarities directly for the multiscale data obtaining in this way a multiscale dissimilarity representation. When these dissimilarities are specifically designed for the problem, it is even possible that they do not fulfill metric requirements. Therefore, standard statistical analysis techniques may not be easily applicable. We propose a new method to combine non-metric multiscale dissimilarities in a compact representation which is used for classification. The method is based on the extended multiscale dissimilarity space and prototype selection, which allows us to handle the potentially non-metric nature of the dissimilarities and exploit the multiscale information at the same time. This is achieved in such a way that the most informative examples per scale are selected. Experimental results show that the approach is promising since it finds a better trade-off in accuracy and efficiency than its counterpart approaches.

Keywords: Extended multiscale dissimilarity space · Multiscale data · Prototype selection · Genetic algorithms

1 Introduction

The term multiscale refers to data represented at different scales of resolution. If the multiscale data is provided by an expert in the form of dissimilarities, the

Parts of the work described have been published as part of the first author's Ph.D. Thesis [1].

Yan Li is now with Lely Technologies, Maassluis, The Netherlands.

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 150–157, 2017.
DOI: 10.1007/978-3-319-52277-7_19

following question remains open: how to properly use the information contained in multiscale dissimilarities for classification? In the literature on supervised pattern recognition for multiscale similarities, we can find two main approaches: scale selection [2], and scale combining [3,4]. Scale selection has been tackled, for example, by Multiple Kernel Learning (MKL) [2], which is similar to the problem of selecting the best kernels for a given problem. For scale combining, all the different scales may be combined in the form of similarity or kernel matrices using MKL as well. However, note that, for potentially non-metric data, it is not possible to use kernel methods in a straightforward manner. One option is creating kernels from the dissimilarities, and after an eigen-analysis, correcting eigenvalues of the non-metric matrix by applying a strategy as spectrum clipping etc [5], which leads to some information loss.

There are other approaches that deal with a dissimilarity matrix which is potentially non-metric; for instance, the k Nearest Neighbour classifiers (k-NN) directly applied to the matrix, and the classifiers in the Dissimilarity Space (DS) [6,7]. In both cases, classifiers can be trained on the individual scales and combination may be performed by standard classifier combiners. Another possibility to combine the scales is by computing a weighted average of the dissimilarities [4]. The disadvantage of this approach is its high computational cost since, for an incoming test object, the dissimilarities with all the prototypes that span the DS for all the scales must be computed. Another approach, to which little attention has been paid so far, is constructing an Extended Multiscale Dissimilarity Space (EMDS) from the dissimilarity matrices [8]. Despite the fact that the first results presented in [8] using the EMDS were discouraging, we consider that a smarter selection of the set of prototypes can lead to better results.

In this paper we propose the use of a selection criterion in the EMDS optimized by a simple Genetic Algorithm [9] to perform such a smart selection. This criterion preserves the most important information from all the scales using the most informative prototypes. In our approach, a smart compromise is obtained between scale selection and scale combination, avoiding also expensive methods such as classifier combination.

The remaining part of the paper is organized as follows. Section 2 introduces the EMDS, presents the related work on prototype selection and the description of the proposed strategy to reduce the EMDS. Section 3 presents the data, experimental setup, results and analysis. Conclusions are drawn in Sect. 4.

2 Proposed Method

The DS was proposed by Pekalska and Duin [6] as an alternative to represent dissimilarity data. The DS is an adequate option to handle measures that are non-Euclidean or even non-metric. All the statistical pattern recognition procedures suitable for Euclidean spaces can be applied to the DS. Let X be the space of objects into consideration which may not be a feature vector space but a non-standard one such as a graph space. A set of prototypes $R = \{r_1, r_2, \ldots, r_l\} \in X$, also called representation set, is used for the creation of the DS. A training set

$T = \{x_1, x_2, \ldots, x_n\} \in X$ is represented in the DS by the dissimilarities of objects in T with objects in R. In general, for a representation set of l prototypes, and a suitable dissimilarity measure for the problem $d : X \times X \rightarrow \mathbb{R}_0^+$, we obtain a dissimilarity matrix $D(T, R)$; the mapping to a DS is represented as $\phi_R^d : X \rightarrow \mathbb{R}^l$. The representation of an object x in the DS is the vector of its dissimilarities with the prototypes: $\phi_R^d(x) = [d(x, r_1)\ d(x, r_2)\ \ldots\ d(x, r_l)]$.

The extended space representation is created from the individual representations in a DS for each scale. For a multiscale problem with M scales, denoting $D_m = D_m(T, R)$ the dissimilarity matrix computed for scale m, we have $D_1, D_2, \ldots, D_M$, normalized dissimilarity matrices. The representation of training objects in the EMDS is created by the concatenation of the individual dissimilarity matrices for each scale: $[D_1\ D_2\ \ldots\ D_M]$. The embedding of any object is obtained by the mapping $\Theta_R^d : X \rightarrow \mathbb{R}^{lM}$, which returns the vector of the dissimilarities with the prototypes from all the scales:

$$\Theta_R^d(x) = [\phi_{R_1}^d(x^1)\ \phi_{R_2}^d(x^2) \ldots \phi_{R_M}^d(x^M)], \tag{1}$$

where $R_m = \{r_1^m, r_2^m, \ldots, r_l^m\} \in X_m, m = 1...M$, is the representation set in scale m and X_m is the space of objects for scale m; $x^m \in X_m, m = 1...M$, are the representations of x under the different scales.

The main problem with the EMDS is its high dimensionality. It is a cause of overfitting and the "curse of dimensionality" phenomenon. Another problem is the increase of the computational costs involved in classification. In order to be able to use the multiscale information avoiding these problems, a prototype selection must be performed to create a reduced EMDS. As the EMDS presents different conditions compared to a standard DS, it is not possible to use most of the prototype selection procedures such as the KCentres or ModeSeek [7] unless they are applied on a single scale as in [8]. These methods require a direct comparison of the prototypes being analyzed, but for the EMDS these prototypes are not directly comparable since they belong to different scales. Another good method, the Forward Selection (FS) [7], is not adequate for the EMDS due to the high dimensionality of this space and the method being quadratic in complexity with respect to this dimensionality. However, we found that GAs are more adequate for dealing with the EMDS, therefore we focus on this optimization strategy to select the prototypes in the EMDS.

We consider that GAs are specially suitable for prototype selection in dissimilarity representations, since, similar or nearby objects carry a similar informational value and they can be chosen indistinctively as prototypes. Therefore, in our prototype selection problem, a thorough search is not needed, and there may be many suboptimal solutions that are sufficiently good and very close to the optimal one. Due to this, a GA can find good solutions for the prototype selection problem very fast, in contrast to other applications where GAs may converge slowly. The GA for prototype selection in a DS was proposed in [9], where it showed a good performance in standard DSs of moderate dimensionality. However, its performance for very high dimensional spaces such as the extended ones has not been studied.

Our criterion focuses on selecting the prototypes in the EMDS taking into account the information provided by all the scales simultaneously, and not by a single scale as in previous approaches. The proposed criterion counts the matching labels of the prototypes and their nearest objects in each scale. For a given cardinality, the winning set of prototypes is the one that maximizes this number, note that selected prototypes for one scale are not necessarily selected for other scales. Only the combination (prototypes; scale) with significant contribution to the maximization is selected. This criterion can be formulated as follows: $\jmath = \sum_{x_i \in S} ML(x_i)$, where:

$$ML(x_i) = \begin{cases} 1, \lambda_S(x_i) = \lambda_S(x_k) \\ 0, \lambda_S(x_i) \neq \lambda_S(x_k) \end{cases}, x_k = \underset{x_j \in S \setminus \{x_i\}}{\operatorname{argmin}} \; d(x_i, x_j) \tag{2}$$

where $\lambda_S(x_i)$ and $\lambda_S(x_k)$ are the class labels of x_i, x_k respectively, and x_k is the object with minimum Euclidean distance to x_i in the DS. The criterion $\jmath$ is therefore the number of matching labels for the candidates set S in the dissimilarity space.

The GA is an evolutionary method which uses heuristics to converge to better solutions, resembling biological processes such as crossover and mutation. Each potential solution (individual, chromosome) is a set of prototypes of fixed cardinality l codified in a $l - tuple$ of prototypes indexes. Note that, by resorting to this type of solution representation, the parameter that influences the memory requirements is the desired number of prototypes (usually small $l \in [10, 100]$) and not the dimensionality of the EMDS. The GA starts the search in an initial population of individuals randomly generated. In each evolution cycle, it evaluates the population using the fitness function. The population undergoes crossover (with best fitted individuals) and mutation processes until the criteria are met. In our approach, we use uniform crossover since each chromosome reproduces with the best fitted one with a preset probability, usually 0.5, and elitist selection since the best fitted individual is retained for the next generation without undergoing mutation. Besides, the population minus the best fitted individual undergoes mutation with a small preset probability, e.g. 0.05. The probabilities for mutation and crossover are usually set in a way that a good trade-off between "exploitation versus exploration" is obtained. The exploitation means the GA searches in a local region of the space where the last good solution was found (by setting crossover probability), and by the exploration the GA searches in a larger and more global region of the space (by setting mutation probability) to avoid loosing good solutions that may not be in a local neighbourhood of the last good solution.

3 Experimental Analysis

In this section, the multiscale data sets used in our experiments are described. The different approaches for prototype selection are presented. The experimental setup, results and discussion are also provided.

Data. It is worth noting that it is very difficult to collect real-world multiscale data sets where the dissimilarities were proposed by experts as suitable for the problem and where the multiscale information made sense according to them. Three different multiscale data sets were collected for the experiments. They are the Colon, Texture and Chicken Pieces data sets. Their descriptions are as follows. The *Colon* data set represents colon tissue data; it was provided by Dr. Marius Nap from the Atrium Medical Center in Heerlen, The Netherlands. The objects are microscope image patches of size 1024 × 1024 belonging to four classes: normal, inflamed, adenomatous, and cancer. The Laplacian of different scales was applied to each image patch, and the city-block (L1) distance between the histograms of the response images was used as the dissimilarity measure. The *Texture* data set (Brodatz) was downloaded from [10]. It has 111 images that we consider as classes. The 640 × 640 images were partitioned into 9 subimages that are used as class objects. The Leung-Malik filter set at different scales was applied to the images, and the Chi square distance between the histograms of the response images was computed. The *Chicken Pieces* data set [11] contains images in binary format representing silhouettes from five different parts of the chicken. From these images the edges are extracted and approximated by segments of different pixel length, and string representations of the angles between the segments are derived. The dissimilarity matrix is composed by edit distances between these strings. A description of the data sets is presented in Table 1.

Table 1. Properties of the multiscale data sets, the last column ($|T|$) refers to the training set cardinality used for the experiments

Data sets	# Classes	# Obj	# scales	$\|T\|$ in EMDS
Colon	4	375×4	9	100×9
Texture	111	9×111	6	222×6
Chicken pieces	5	446	11	170×11

Experimental Setup. In the experiments our aim is to show that the reduced EMDS obtained by the prototype selection may provide a more compact and discriminative representation of multiscale dissimilarities compared to the space of averaged multiscale dissimilarities, and the DS created by the best individual scale. Note that in the comparison we always use the same dimensionality of the spaces and the same length of vectors codifying the data. For consistency, we compare the same Linear Discriminant classifier (LDC), which is the Bayes classifier assuming normal densities with identical covariance matrices, and the 1-NN in the different spaces and data sets. All the dissimilarity matrices were normalized to avoid scaling problems by setting the mean dissimilarity per scale to 1 using global rescaling.

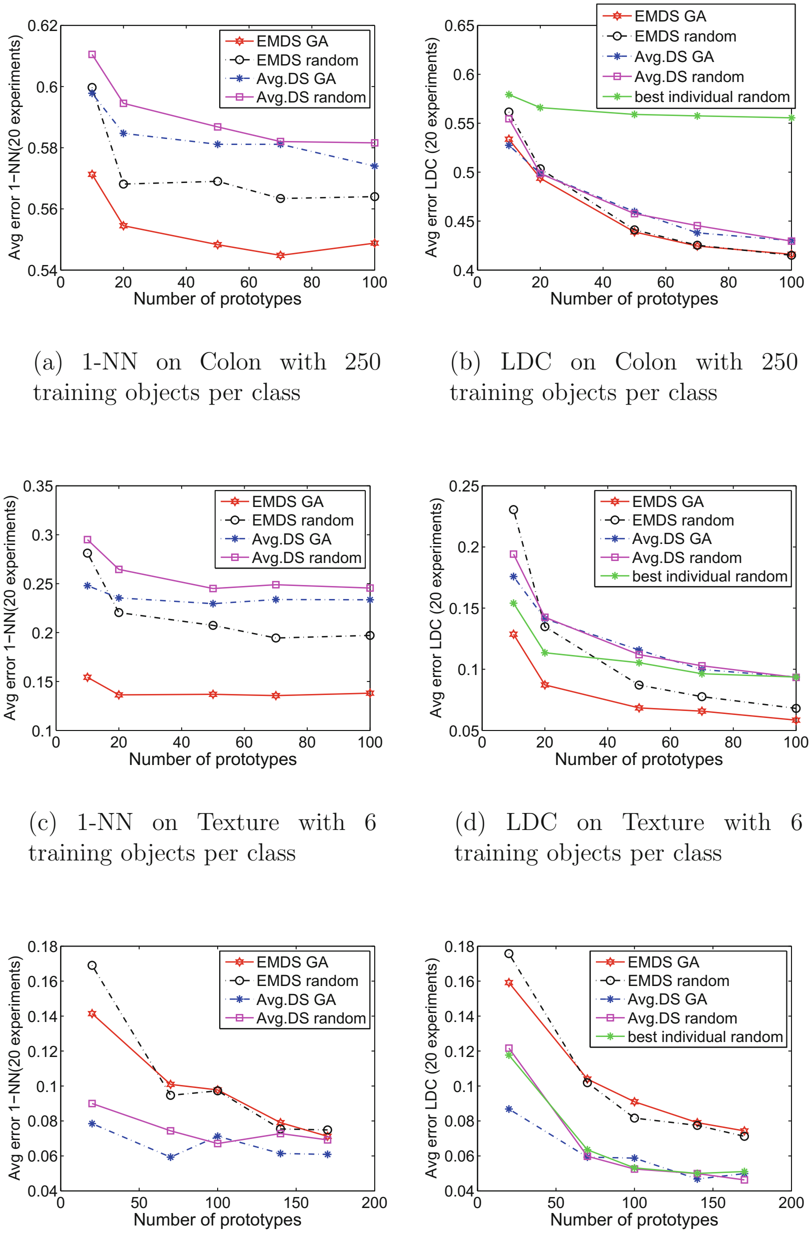
Since the data sets present a small size, they were 20 times randomly divided into two sets: a training set, that was used as the candidate set of prototypes to optimize the selection criterion and to build the final classifiers in the EMDS;

and a test set, which was only used to compute the final classification error. The prototype selectors executed are: GA in the EMDS (our proposal), random selection in the EMDS (as a baseline for the GA) and recent approaches from the literature on combining non-metric multiscale dissimilarities, also reducing the set of prototypes to ensure comparability of results: GA in the averaged DS, random selection in the averaged DS, random selection in the individual DS for each scale (only for the best performing classifier since for the other one a similar behaviour was found). Different parameters have been proposed for the GA, however, they can be problem-dependent. Thereby we decided to use parameters proposed in previous works on prototype selection [1]: Initial population: 30 individuals or solutions, Probability of reproduction: 0.5, Probability of mutation: 0.05, Stopping condition: 20 generations reached. We analyzed these parameters for our problem and we found that the GA converged to good solutions after 10 generations, but setting 20 generations as stopping condition ensured slightly better results. The results are stable in general for small variations of the parameters, but not for large variations.

Results and Discussion. Figure 1 present the results obtained for the data sets used in our study. Classification errors are presented for increasing numbers of prototypes in multiscale spaces as well as in the individual spaces from the different scales. Standard deviations were not included to maintain the clarity of the plots, but they vary between 0.02 and 0.05 for Chicken Pieces, 0.01 and 0.03 for Colon, and between 0.007 and 0.05 for Texture data set.

For the Colon data set, it can be seen that the EMDS outperforms the averaged DS and the individual scales. Results for the Texture data set show a clear example where the EMDS provides better results than those of the other approaches. In this data set as well as in the Colon data set, the EMDS significantly improves the results of the individual scales. Results for the Chicken Pieces data set show a different behaviour. The averaged DS outperforms the EMDS. We believe that this happens because, in this data set, only four scales present a low classification performance while seven scales perform similarly well. These large number of good performing scales influence the average dissimilarity computation heavier than the four bad ones. However, for the Colon and Texture data sets, the individual scales perform significantly different from each other, and the averaged space suffers from this while the reduced EMDS is able to capture the complementary information for classification. It can also be seen that the proposed selection outperforms the random selection, which usually performs very good for high dimensional DSs.

Note that the proposed approach is less computationally expensive than the combination of the scales by averaging all the DS representations. In the averaging case, the dissimilarities with the prototypes in all the scales must be computed while in our approach only the dissimilarities with the prototypes in the specific scale they were selected are computed. Therefore, for N scales and p prototypes and z number of objects, our approach computes $z \times p$ dissimilarities, while the average approach computes $z \times p \times N$ dissimilarities.



(a) 1-NN on Colon with 250 training objects per class

(b) LDC on Colon with 250 training objects per class

(c) 1-NN on Texture with 6 training objects per class

(d) LDC on Texture with 6 training objects per class

(e) 1-NN on Chicken Pieces with 350 training objects per class

(f) LDC on Chicken Pieces with 350 training objects per class

Fig. 1. Classification results on the extended (EMDS), averaged (avg. DS) and individual dissimilarity spaces for the different data sets

4 Conclusions

In this paper, we proposed a new strategy to represent potentially non-metric multiscale dissimilarity data in a compact and discriminative way. The multiscale representation is achieved by the extended dissimilarity space while the compact representation is achieved by means of a selection criterion optimized by a GA in a way that the most informative examples per scale are selected.

The classification results using the proposed compact multiscale representation outperform results using the representations in individual scales, despite having the same computational cost. In addition, the proposed approach is less computationally expensive than the combination of the scales by averaging all the DS representations, even improving the classification accuracies when the individual scales provide complementary information. Future work will focus on better characterizing the data sets where this approach is useful.

References

1. Plasencia-Calaña, Y.: Prototype selection for classification in standard and generalized dissimilarity spaces. Ph.D. thesis, Delft University of Technology, September 2015
2. Gönen, M., Alpaydin, E.: Multiple kernel learning algorithms. J. Mach. Learn. Res. **12**, 2211–2268 (2011)
3. Liu, Y.M., Ye, L.B., Zheng, P.Y., Shi, X.R., Hu, B., Liang, J.: Multiscale classification and its application to process monitoring. J. Zhejiang Univ. Sci. C **11**, 425–434 (2010)
4. Li, Y., Duin, R.P.W., Loog, M.: Combining multi-scale dissimilarities for image classification. In: Proceedings of the 21th International Conference on Pattern Recognition, ICPR 2012. IEEE Computer Society (2012)
5. Gisbrecht, A., Schleif, F.M.: Metric and non-metric proximity transformations at linear costs. Neurocomputing **167**, 643–657 (2015)
6. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co. Inc., River Edge (2005)
7. Pekalska, E., Duin, R.P.W., Paclík, P.: Prototype selection for dissimilarity-based classifiers. Pattern Recogn. **39**(2), 189–208 (2006)
8. Ibba, A., Duin, R.P.W.: A multiscale approach in combining classifiers in dissimilarity representations. In: Gevers, T., Bos, H., Wolters, L. (eds.) 15th Annual Conference of the Advanced School for Computing and Imaging, ASCI 2009 (2009)
9. Plasencia-Calaña, Y., García-Reyes, E., Orozco-Alzate, M., Duin, R.P.W.: Prototype selection for dissimilarity representation by a genetic algorithm. In: Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR 2010, pp. 177–180. IEEE Computer Society, Washington (2010)
10. Randen, T.: Brodatz textures. http://www.ux.uis.no/~tranden/brodatz.html
11. Bunke, H., Buhler, U.: Applications of approximate string matching to 2D shape recognition. Pattern Recogn. **26**(12), 1797–1812 (1993)