Bio-Chemical Data Classification by Dissimilarity Representation and Template Selection

Victor Mendiola-Lau[1], Francisco José Silva Mata[1(✉)], Yenisel Plasencia Calaña[1], Isneri Talavera Bustamante[1], and Maria de Marsico[2]

[1] Advanced Technologies Application Center, Havana, Cuba
{vmendiola,fjsilva,yplasencia,italavera}@cenatav.co.cu
[2] Universita degli Studi de Roma, Rome, Italy
demarsico@di.uniroma1.it

Abstract. The identification and classification of bio-chemical substances are very important tasks in chemical, biological and forensic analysis. In this work we present a new strategy to improve the accuracy of the supervised classification of this type of data obtained from different analytical techniques that combine two processes: first, a dissimilarity representation of data and second, the selection of templates for the refinement of the representative samples in each class set.

In order to evaluate the performance of our proposal, a comparative study between three approaches is presented. As a baseline, entropy template selection (ETS) is performed in the original feature space and selected templates are used for training. The underlying concept of the other two alternatives, is the combination of Dissimilarity Representations and ETS. The first alternative performs ETS in the original feature space and uses the selected templates as prototypes for the generation of the dissimilarity space and as training set. The second one represents the data in the dissimilarity space, and next ETS is performed.

The experimental results showed that an adequate combination of the representation in the dissimilarity the space and the selection of templates based on entropy, outperformed the baseline in accuracy and/or efficiency for the majority of the problems studied.

Keywords: Dissimilarity representation · Entropy
Template selection · Classification · Bio-chemical data

1 Introduction

Classification is one of the most common tasks of pattern recognition. It is based on learning the structure of groups of objects with similar patterns. In this learning process, it is assumed that there are training examples that are representative and contain sufficient information to find a good (generalizable) model for predefined groups/classes of those examples. Therefore, new unseen

© Springer International Publishing AG, part of Springer Nature 2018
M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 374–381, 2018.
https://doi.org/10.1007/978-3-319-75193-1_45

objects can be assigned into these groups reliably. The formalized representation of objects is an important aspect in the determination of a good class description.

The identification and classification of bio-chemical substances are very important tasks in chemical, biological and forensic analysis. Nowadays, bio-chemical data are analyzed as vectors of discretized data where the variables have no connection, and other aspects of their functional nature shape differences, are also ignored. The goal is to find new data representations with the capability to extract more information from the data taking into account its specific characteristics, which should improve classification results.

Dissimilarity Representation (DR) is rather a new approach which can take the functional information into account, and has shown to be advantageous in problems where the number of objects is small, and also when they are represented in high dimensional spaces, which are both common characteristics of bio-chemical spectral data sets.

Profound studies of the dissimilarity on chemical and bio-chemical data sets have not been carried out. There are already some results on spectral data in general [5,14], demonstrating its advantages for classification but a generalization of this behavior for the diversity of spectral and non-spectral analysis techniques and the selection of dissimilarity measures more appropriate for the typicity of the data has not yet been sufficiently researched. In consequence, there is still no consensus on the real scope of the improvements that can be obtained with the use of dissimilarity-based representations in a broader set of analytical techniques, underpinned by thorough experimentation and analysis of the results.

Given the wide physical and chemical nature of the data studied, the notion dissimilarities/proximities between samples play a key factor. Another factor to take into account is the representation chosen for building the dissimilarity space [12]. Recent works [4,10] have addressed this problem by selecting templates based on an entropy criterion. A suitable combination of these two factors is key for the improvement of the classification process. In this work, some approaches for combining dissimilarity representations with an entropy-based template selection are proposed (see Fig. 1):

- *(i) ENTR+CLAS* (a) perform ETS in the original feature space and the selected templates used as a training set (baseline).
- *(ii) ENTR+DR+CLAS* (b) perform ETS as in *(i)*, the selected templates are used as prototypes to build a dissimilarity space in which they are also used as a training set for classifiers.
- *(iii) DR+ENTR+CLAS* (c) build a dissimilarity space and in which to perform ETS and classifier validation.

2 Materials and Methods

The data sets used in this study come from different analytical techniques and sources, thus allowing for representativeness. Some data sets such as *IR Fuel* (Infrared spectroscopy of fuel) [13], *GC Cannabis* (Gas Chromatography of cannabis) [10], and *UV Cannabis* (Ultraviolet spectroscopy of cannabis) [10]

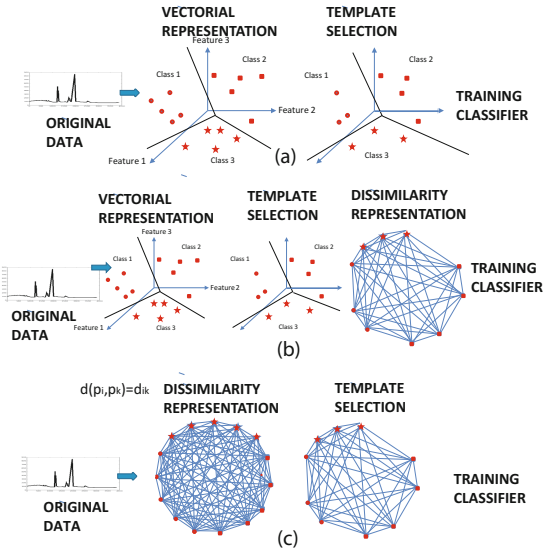


Fig. 1. Approaches for combining dissimilarity representations with entropy.

were obtained from our laboratory. *MVDA Alcohol* (Alcohols in medical applications) [8], *NIR Tablets* (NIR spectroscopy of tablets) [1], *NIR Tecator* (NIR spectroscopy of meat) [16], *Raman Porkfat* (Raman spectroscopy of porcine tissue) [2] and *Raman Tablets* (Raman spectroscopy of tablets) [1]; come from known sources. Due to the variety of the different analytical techniques, traditional data preprocessing methods like Multiplicative Scatter Correction (MSC) [7] and Savitzky-Golay (SG) derivatisation [15] were used.

3 Dissimilarity Representations

In the traditional framework of Pattern Recognition, feature representations are usually employed for objects. The Dissimilarity Representation [6], consists of an alternative representation for objects, where proximities and/or dissimilarities among them play a key role. Let X a set of objects in some space (which might not be a vector space), and $T = \{x_1, x_2, \ldots, x_N\}$ a finite sample of such space ($T \subset X$) consisting of N samples. Let $d : X \times X \rightarrow \mathbb{R}^+$ be a dissimilarity measure denoting the notion of resemblance between objects in T. A dissimilarity representation of an object x in T, denoted as $D(x,T) = [d(x,p_1), d(x,p_2), \cdots, d(x,p_N)]$, is a vector containing the dissimilarities between x and the objects in T. While building a dissimilarity representation, a representation set $R = \{r_1, r_2, \ldots, r_k\}$ ($R \in X$), is often employed. Using this representation set, an object x can be represented as $D(x,R) = [d(x,r_1), d(x,r_2), \cdots, d(x,r_k)]$, a vector containing the dissimilarities between x and each *prototype* in R. Each new *feature* corresponds to the dissimilarity with some prototype and the dimension of the space is determined by the amount of prototypes used.

The Dissimilarity Representation requires a suitable measure that expresses the relation among objects, in this case between the spectra. For substances, the concentration of its components and their presence is a key factor for a proper comparison. Therefore, the dissimilarity measures employed should take this relation into account, which different for every kind of spectra or data content [9]. In such manner, a more powerful representation can be obtained, allowing for a better discrimination.

4 Entropy-Based Template Selection

In previous works [4,10], the information gain has been used to select the best templates for improving classification accuracy. In analytic chemistry, the analysis performed on a substance is usually replicated, yielding almost identical replicated samples. The entropy-based template selection strategy proposed in [4] allows to identify the most representative templates in a given set. Thus, it is possible to reduce the size of the training set for classifiers by removing less significant samples. This procedures is outlined next.

Let G be a set of templates that can be expressed as the union of several subsets (groups), i.e., $G = \cup_{k=1}^{K} G_k$ and $G_k \cap G_h = \varnothing$, $\forall k \neq h$. Let $s : G \times G \rightarrow \mathbb{R}^+$ be a similarity measure defined over templates in G and the comparison of templates v and $g_{i,k} \in G_k$ be denoted as $s_{i,v} = s(v, g_{i,k})$, which is normalized to be a value in the interval [0, 1] [4]. Considering a distance/dissimilarity measure $d_{i,v} = d(v, g_{i,k})$, it is possible to obtain a similarity measure $s_{i,v} = \frac{1}{1+d(v,g_{i,k})}$, which has the advantages of being normalized in [0, 1] and avoid numeric errors for very small distance/dissimilarity value. Assuming that template v was correctly assigned to the class k, each $s_{i,v}$ value can be interpreted as the probability that template v conforms to $g_{i,k}$, that is $s_{i,v} = p(v \approx g_{i,k})$.

The entropy of the probability distribution obtained by applying (4) to the whole G_k with respect to a probe v is defined as follows:

$$H(G_k, v) = -\frac{1}{\log_2 |G_k|} \sum_{i=1}^{|G_k|} s_{i,v} \log_2 (s_{i,v}), \tag{1}$$

where $\frac{1}{log_2(|G_k|)}$ is a normalization factor. Finally, the entropy value for G_k is computed by considering each template $g_{j,t} \in G_k$ as a probe v:

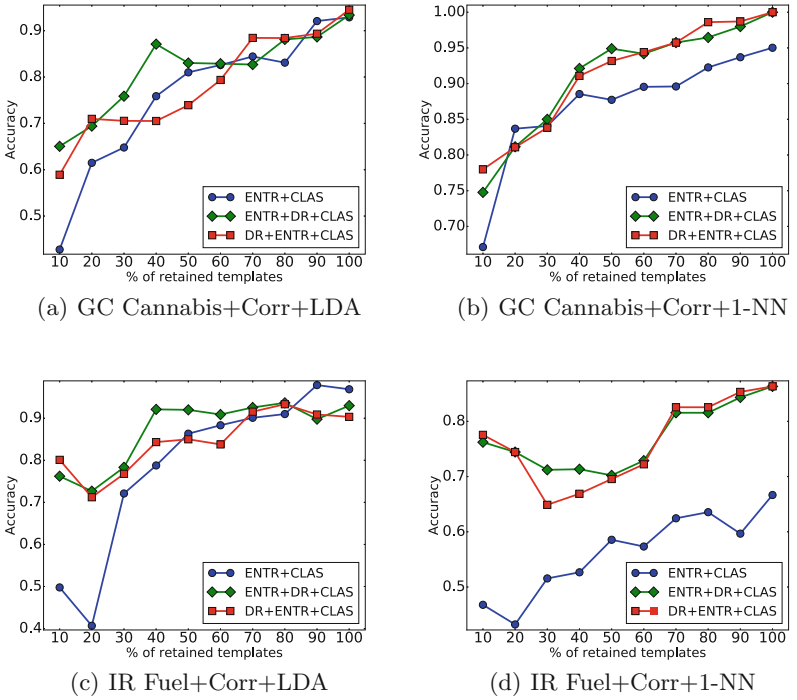
$$H(G_k) = -\frac{1}{\log_2 |Q|} \sum_{q_{i,j} \in Q} s_{i,j} \log_2 (s_{i,j}), \tag{2}$$

where Q represents the set of pairs $q_{i,j} = (q_{i,k}, q_{j,k})$ of elements in G_k such that $s_{i,j} > 0$. $H(G_k)$ represents a measure of heterogeneity for G_k, and it can be used to select a subset of representative samples out of it. The procedure described in [4] is based on an ordering of the gallery templates according to a representativeness criterion.

After performing sample ordering according, a *top-percentage* criterion [10] was used to select templates that guarantee a suitable representativeness of the gallery. In this work, the classification behavior for different training sets yielded by the *top-percentage* criterion was analyzed.

5 Experimental Results and Discussion

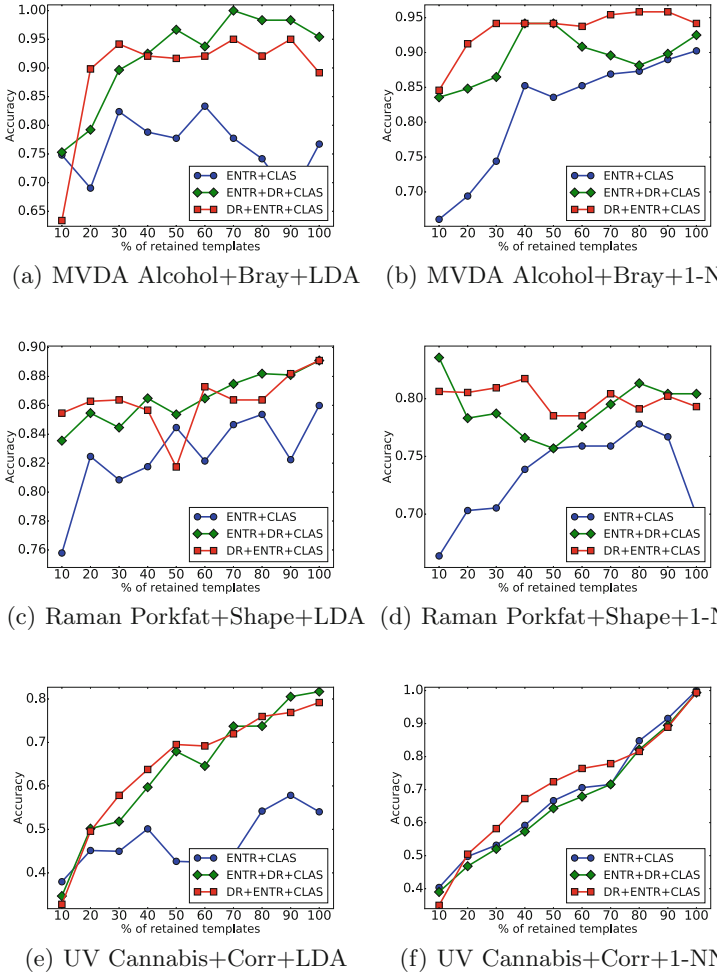
As previously mentioned in Sect. 2, a representative experimental set of 8 spectral and non-spectral data sets were employed for evaluating the proposed strategies. Although tests were carried out in all data sets, the ones showing the general behavior for each analytical technique are discussed. A total of 14 dissimilarity measures were used for data representation in dissimilarity spaces and those achieving the highest classification accuracy for each data set are presented. The dissimilarity measures that achieved the best results were *DShape* (Shape) [11], *Correlation distance* (Corr) [3] and *Bray Curtis* (Bray) [3]. The validation protocol consisted of averaging the results of 5 independent runs of a 10-fold cross validation procedure for each approach: *ENTR+CLAS* (baseline), *ENTR+DR+CLAS* and *DR+ENTR+CLAS*. Traditional classifiers such as LDA and 1-NN were employed for validation.



(a) GC Cannabis+Corr+LDA (b) GC Cannabis+Corr+1-NN

(c) IR Fuel+Corr+LDA (d) IR Fuel+Corr+1-NN

Fig. 2. Classification accuracies for different data sets.



(a) MVDA Alcohol+Bray+LDA (b) MVDA Alcohol+Bray+1-NN

(c) Raman Porkfat+Shape+LDA (d) Raman Porkfat+Shape+1-NN

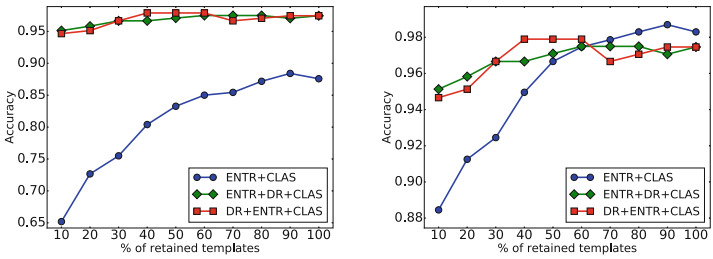
(e) UV Cannabis+Corr+LDA (f) UV Cannabis+Corr+1-NN

Fig. 3. Classification accuracies for different data sets.

Figures 2 and 3 show a substantial improvement in the classification accuracy of our proposals for the majority of data sets. Moreover, the *ENTR+DR+CLAS* strategy managed to obtain a very good accuracy while maintaining a reduced training set size and a lower number of features, thus reducing the time and spatial complexity of the classification process. The dissimilarity representation proved to be advantageous for both classifiers studied. First, the 1-NN in feature spaces is very sensitive to outliers since its decision is only based on the nearest object from the training set to the test object. If this nearest object is an outlier or is noisy, the classification decision will be wrong. However, in the dissimilarity space, the dissimilarity to an object or prototype accounts only for one feature; therefore outlier objects are compensated by the other prototypes which are

not noisy. By compensating the influence of outliers, the wrong decision can be corrected. In addition, the LDA in a high dimensional feature space suffers from the curse of dimensionality and the small sample size problem and usually must be regularized. The dissimilarity space offers an intrinsic regularization for LDA by prototype selection.

Another issue that should not be overlooked is the fact that dissimilarity representations proved to be more robust than the orginal feature space for unprocessed data sets (see Fig. 4) (also observed in *Raman Tablets* with Savitzky-Golay smoothing). In preprocessed data sets, the 1-NN classifier showed improvements for the baseline strategy. The 1-NN classifier heavily relies in distances computation in the original space, and therefore, its accuracy increased for preprocessed data sets. On the other hand, the LDA classifier had a more robust behavior for unprocessed data sets.



(a) DShape+1-NN (b) MSC+DShape+1-NN

Fig. 4. MSC preprocessing effect on *NIR Tecator* data set.

An exploratory analysis ought to be conducted in order to determine the best number of templates to be selected for an optimal classification process. On the other hand, as observed in the majority of data sets, a compromise can be made among a near-optimal accuracy and a considerably smaller set for training the classifiers.

6 Conclusions and Future Work

In this work, new approaches for combining dissimilarity representations with an entropy-based template selection strategy were proposed. Our proposals showed an improvement over operating in the original feature space for classification tasks. It is also important to stress that our proposals were also able to deal with the curse of dimensionality present in some of the data sets. Moreover, the proposed strategies can be used for discovering more compact training sets by means of an exploratory analysis, while allowing to preserve, and sometimes improve, the classification accuracy. As could be observed, performing an entropy-based template selection in the feature space and using the selected instances as prototypes for a dissimilarity representation not only achieved very good results,

but also yielded a more compact representation and increased the classification performance. As future work, a more extensive and detailed analysis ought to be conducted concerning the suitability of the existing dissimilarity measures for the different types of spectral data studied. Also, further tests should be carried out on a wider range of classification strategies.

References

1. Quality and Technology website (2017). http://www.models.life.ku.dk/tablets
2. Quality and Technology website (2017). http://www.models.life.ku.dk/RAMANporkfat
3. Cha, S.H.: Comprehensive survey on distance/similarity measures between probability density functions. City **1**(2), 1 (2007)
4. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: Entropy-based template analysis in face biometric identification systems. Sig. Image Video Process. **7**(3), 493–505 (2013)
5. Duin, R.P., Pekalska, E.: The dissimilarity space bridging structural and statistical pattern recognition. Pattern Recogn. Lett. **33**(7), 826–832 (2012)
6. Duin, R.P., et al.: The dissimilarity representation for pattern recognition: foundations and applications, vol. 64. World Scientific (2005)
7. Helland, I.S., Næs, T., Isaksson, T.: Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. Chemometr. Intell. Lab. Syst. **29**(2), 233–241 (1995)
8. Infometrix: Infometrix website (2017). https://infometrix.com/pirouette/
9. Kumar, V., Chhabra, J.K., Kumar, D.: Performance evaluation of distance metrics in the clustering algorithms. J. Comput. Sci. **13**(1), 38–52 (2014)
10. Mendiola-Lau, V., Mata, F.J.S., Martínez-Díaz, Y., Bustamante, I.T., de Marsico, M.: Automatic classification of herbal substances enhanced with an entropy criterion. In: Beltrán-Castañón, C., Nyström, I., Famili, F. (eds.) CIARP 2016. LNCS, vol. 10125, pp. 233–240. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52277-7_29
11. Paclik, P., Duin, R.: Classifying spectral data using relational representation. NA (2003)
12. Plasencia Calaña, Y.: Prototype selection for classification in standard and generalized dissimilarity spaces. Ph.D. thesis, TU Delft (2015)
13. Porro Munoz, D.: Classification of continuous multi-way data via dissimilarity representation (2013)
14. Porro-Muñoz, D., Talavera, I., Duin, R.P., Hernández, N., Orozco-Alzate, M.: Dissimilarity representation on functional spectral data for classification. J. Chemom. **25**(9), 476–486 (2011)
15. Press, W.H., Teukolsky, S.A.: Savitzky-Golay smoothing filters. Comput. Phys. **4**(6), 669–672 (1990)
16. Thodberg, H.H.: StatLib–Datasets Archive website (2017). http://lib.stat.cmu.edu/datasets/tecator