Digital Signal Processing 27 (2014) 69–78

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

An empirical study on improving dissimilarity-based classifications using one-shot similarity measure ☆

Sang-Woon Kim

Department of Computer Engineering, Myongji University, Yongin, 449-728, South Korea

ARTICLE INFO

Article history:
Available online 8 January 2014

Keywords:
Statistical pattern recognition
Dissimilarity-based classification (DBC)
Semi-supervised learning (SSL)
One-shot similarity (OSS) measure

ABSTRACT

This paper reports an experimental result obtained by additionally using unlabeled data together with labeled ones to improve the classification accuracy of dissimilarity-based methods, namely, dissimilarity-based classifications (DBC) [25]. In DBC, classifiers among classes are not based on the feature measurements of individual objects, but on a suitable dissimilarity measure among the objects instead. In order to measure the dissimilarity distance between pairwise objects, an approach using the one-shot similarity (OSS) [30] measuring technique instead of the Euclidean distance is investigated in this paper. In DBC using OSS, the unlabeled set can be used to extend the set of prototypes as well as to compute the OSS distance. The experimental results, obtained with artificial and real-life benchmark datasets, demonstrate that designing the classifiers in the OSS dissimilarity matrices instead of expanding the set of prototypes can further improve the classification accuracy in comparison with the traditional Euclidean approach. Moreover, the results demonstrate that the proposed setting does not work with non-Euclidean data.

1. Introduction

The aim of this paper is to report an empirical result obtained by additionally using unlabeled data together with labeled ones to improve the classification accuracy of dissimilarity-based methods, namely, dissimilarity-based classifications (DBC) [25]. In DBC, defining classifiers among the classes is not based on the feature measurements of individual objects, but rather on a suitable dissimilarity measure among the individual objects. The advantage of this strategy is that it offers a different way to include expert knowledge on the objects in classifying them [10]. A few of the issues we encounter when designing DBCs are as follows: selecting (creating) the prototype subset from a given data set [18,21,26]; reducing the dimensionality of the dissimilarity space [16,28]; solving non-Euclidean problems in the dissimilarity space (pseudo-Euclidean embedding) [10]; increasing the robustness of the dissimilarity space (or combining dissimilarity representations) [17]; optimizing classification (or clustering) based on dissimilarity increments (i.e., differentiation of dissimilarity distances) [2,13].

In order to explore the other issues, various strategies have been proposed in the literature. Among them, investigations have focused specifically on generalizing the dissimilarity representation by using various methods, such as feature lines and feature planes [23,24] and hidden Markov models [3]. In [23], the authors enrich (generalize) the dissimilarity representation by using the nearest feature rules. The generalization provided by the feature lines and/or planes covers all the possible intra-class pairs and triplets of prototypes to find the intrinsic geometric information available at the pairwise dissimilarities. The enrichment of the dissimilarity representation is beneficial for a specific structure of data, such as correlated (cigar-like or elongated) datasets having, possibly, a moderately nonlinear structure. In this strategy, however, objects are represented by a vector of dissimilarities with prototype feature lines (or planes) that are computed between objects of the same class. Consequently, the strategy has two drawbacks: the high amount of generated feature lines that increase computational cost [24] and the use of the labels of objects that leads to a supervised learning system.

On the other hand, when designing a DBC with a measuring system, we sometimes suffer from the difficulty of collecting sufficient (labeled) training data for each class. Labeled instances, for example, are often difficult, expensive, or time-consuming to obtain, as they require the services of an experienced expert. Meanwhile, unlabeled data, defined as the samples that do not belong to the classes being learned, may be relatively easy to collect, but the use of this type of data is limited. To address this problem, in a learning framework of *semi-supervised learning* (*SSL*) [1,4,6,29,32,33], a large amount of unlabeled data, together with labeled data, can be utilized to build better classifiers. Because SSL requires less

☆ This work was supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2012R1A1A2041661). A preliminary short version of this paper was partially presented at the ICPRAM2013, the International Conference on Pattern Recognition Applications and Methods, Barcelona, Spain, in February 15–18, 2013.

E-mail address: kimsw@mju.ac.kr.

1051-2004/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.dsp.2013.12.013

human effort and results in higher accuracy, it is of great interest both in theory and in practice [14,19,20].

In DBC, the SSL strategy can also be considered to improve the classification performance. One of the easiest ways with which unlabeled data contribute to learn DBC classifiers is to simply append them to the representation set. Assume that the cardinalities of a training set, T, and the prototype subset (representation set), P, are denoted by $|T|$ and $|P|$, respectively. When employing an additional unlabeled data U (where the sample size of the set U is $|U|$), the cardinality and the dimensionality of the dissimilarity row vectors that result are $|T|$ and $|P| + |U|$, respectively. Here, a prototype selection method can be utilized to reduce the dimensionality of the dissimilarity space. Consequently, in the traditional feature-based classification (FBC), employing SSL strategy leads to increasing the cardinality of the training data, while, in DBC, utilizing the above strategy results in increasing the dimensionality of the training data. However, as in FBC, it is not also guaranteed that increasing the dimensionality leads to a situation in which the classification accuracy is improved.

In order to improve the classification performance of DBC in an SSL fashion, in this paper we use the well-known *one-shot similarity* (*OSS*) [30,31] measuring scheme based on the background information of available extra (unlabeled) data. To achieve this improvement, we first compute the confidence levels of the training data with the OSS distance. We then construct the dissimilarity matrices, where the dissimilarity is measured with the averaged OSS confidence levels. In OSS, when given two vectors, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and an additionally available (unlabeled) data set, A, a measure of the (dis)similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is computed as follows. First, a discriminative model is learned with $\boldsymbol{x}_i$ as a single positive example and A as a set of negative examples. This model is then used to classify the other vector $\boldsymbol{x}_j$, and to obtain a confidence score. Next, a second such score is obtained by repeating the same process with the roles of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ switched. Finally, the (dis)similarity of the two vectors can be obtained by averaging the above two scores.

The major task of this study is to deal with how the dissimilarity distance can be effectively measured. However, when a limited number of objects are available or the representational capability is insufficient to cover the possible variations of data, it is difficult to achieve the desired classification performance in the dissimilarity representation. To overcome this limitation and thereby improve the classification performance of DBC, in this paper we study a way of exploiting additionally available unlabeled data when measuring the dissimilarity distance with the OSS distance. As in SSL for FBC, we use the easily collected unlabeled data as the background data set, A, with which we can enrich the representational capability of the dissimilarity measures. That is, our goal is to effectively measure the dissimilarity distance with the additional unlabeled data as well as the labeled ones. In DBC, the SSL process is realized in *representation* stage, while, in FBC, it is implemented in *generalization* stage.

The main contribution of this paper is to demonstrate that the classification accuracy of DBC can be improved by using the OSS measuring technique based on unlabeled data. More specifically, experiments have been performed to demonstrate that the OSS distance measure performs better than the Euclidean distance measure. Here, the additional unlabeled set is used as well, but now differently than for building the set of prototypes: it is used in the distance measure and not in building the dissimilarity space. The remainder of the paper is organized as follows. In Section 2, after providing a brief introduction to DBC and OSS, we present an explanation for the use of OSS in DBC and an SSL-type DBC algorithm. Following this, in Section 3, we present an experimental setup for the traditional DBC and proposed DBC algorithms. In Section 4, we present the experimental results of artificial and real-life datasets. Finally, in Section 5, we present our concluding remarks as well as some feature works that deserve further study.

2. Related work

In this section, we briefly review the dissimilarity-based classification (DBC) approach and the one-shot similarity (OSS) measure, which are closely related to the present empirical study. The details of these algorithms can be found in the related literature [25,30,31].

2.1. Foundations of DBC [25]

A dissimilarity representation of a set of objects, $T = \{\boldsymbol{x}_i\}_{i=1}^{n} \subset \mathbb{R}^d$ (d-dimensional examples), is based on pair-wise comparisons, and is expressed, for example, as an $n \times m$ dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, where $P = \{\boldsymbol{p}_j\}_{j=1}^{m} \subset \mathbb{R}^d$, a prototype subset, is extracted from T. The subscripts of D represent the set of elements, on which the dissimilarities are evaluated. Thus, each row, $D_{T,P}[i,j]$, corresponds to the dissimilarity between the pairs of objects, $\langle \boldsymbol{x}_i, \boldsymbol{p}_j \rangle$, where $\boldsymbol{x}_i \in T$ and $\boldsymbol{p}_j \in P$. Consequently, when given a distance measure between two objects, $\rho(\cdot,\cdot)$, an object, $\boldsymbol{x}_i$ ($1 \leqslant i \leqslant n$), is represented as a new feature vector, $\delta(\boldsymbol{x}_i, P)$, as follows:

$$\delta(\boldsymbol{x}_i, P) = \left[\rho(\boldsymbol{x}_i, \boldsymbol{p}_1), \rho(\boldsymbol{x}_i, \boldsymbol{p}_2), \ldots, \rho(\boldsymbol{x}_i, \boldsymbol{p}_m)\right]. \tag{1}$$

Here, the generated dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, defines vectors in a *dissimilarity space*, on which the d-dimensional object, $\boldsymbol{x}$, given in the original feature space, is represented as an m-dimensional vector, $\delta(\boldsymbol{x}, P)$ or shortly $\delta(\boldsymbol{x})$. Thus, for a test sample, $\boldsymbol{z}$, we can achieve the classification by invoking a classifier built in the dissimilarity space and operating it on the m-dimensional vector $\delta(\boldsymbol{z})$.

As mentioned previously, the dissimilarity approach is originally developed for objects, not for feature vectors. However, the dissimilarities are now used as features and can be replaced without any problem by similarities. Thus, it should be noted that an approach defined for arbitrary distances between full objects is used for distances measured in a feature space.[1]

On the basis of what we have just explained briefly, a conventional algorithm for DBC is summarized as follows:

1. Select the prototype subset, P, from the training set, T, by using one of the prototype selection methods described in the literature [26].
2. Using Eq. (1), compute the dissimilarity matrix, $D_{T,P}[\cdot,\cdot]$, in which the dissimilarity distance is computed on the basis of the given measure $\rho(\cdot,\cdot)$, such as the Euclidean distance (l_2 norm).
3. For a testing sample, $\boldsymbol{z}$, compute the corresponding row vector, $\delta(\boldsymbol{z})$, by using the same prototype subset and the distance measure used in Step 2.
4. Achieve the classification by invoking a classifier built in the dissimilarity space and operating it on the vector $\delta(\boldsymbol{z})$.

Here, we can see that the classification performance of DBC relies heavily on how well the dissimilarity space, which is determined by the dissimilarity matrix, is constructed. Thus, to improve the performance, we need to ensure that the dissimilarity matrix is well assembled.

[1] We are grateful to the anonymous referee for providing us with the insight into this.