21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

A study on semi-supervised dissimilarity representation

Cuong V. Dinh [*], Robert P.W. Duin, Marco Loog
Pattern Recognition Laboratory
Delft University of Technology, the Netherlands
{v.c.dinh, r.p.w.duin, m.loog}@tudelft.nl

Abstract

In the dissimilarity representation approach, objects are represented by their dissimilarities with respect to a representation set, rather than by features. Up to now, the representation or prototype set has usually been selected from the training data, limiting the different aspects that can be captured, especially when the training data set is small. This paper studies the performance change if the object's representation is extended by including also test data into the representation set in a semi-supervised setting. Experiments on a set of standard data show that the semi-supervised setting can substantially improve the performance of the dissimilarity based representation especially for the small sample size problem.

1. Introduction

The dissimilarity representation is an approach in which objects are represented by their dissimilarities with respect to others in a data set. It is based on the idea that a class is constituted by objects having similar characteristics. The dissimilarity is small between objects of the same class and large between objects from different classes. Therefore, dissimilarities can be used as discriminant features for classification [5, 3]. The key advantage of the dissimilarity representation is that it provides a way to embed knowledge about the data structural information into powerful feature-based statistical approaches which are intensively available in machine learning and pattern recognition [3].

In the dissimilarity representation approach, the representation set is a set of objects, often called prototypes, to which other objects in the data set are compared. Based on the representation set, a dissimilarity space is constructed in which each dimension corresponds to the distances of all objects to a prototype. The representation set is traditionally selected as the whole training data set or a part of it. When the training set is sufficient large, selecting a small representation set is of interest since it can reduce the computational cost to compute the dissimilarity matrix. In addition, it is shown in [5] that classifiers, such as the quadratic discriminant classifier, when used with prototype selection usually perform better than the 1-NN classifier using the whole training data set.

The situation, however, is different if the training data set is small. The representation set selected from a small training set might miss important prototypes. Consequently, it may limit the different aspects that can be captured in the data and result in an inadequate performance.

In this paper, we aim at finding an improved data representation for the small sample size problem. A nice property of the dissimilarity representation is that it does not necessitate the availability of the labels of the objects in the representation set. In our approach, we enrich the representation set by including also unlabeled samples from the test set in a semi-supervised setting. The assumption we made is that the test set is available during the training process. Our experiments on several standard data sets demonstrate that including unlabeled samples into the representation set often improves the classification results especially for small training size. The semi-supervised dissimilarity representation is therefore useful for applications in which to obtain unlabeled data is much easier than labeled data, e.g. text classification and natural language processing.

2. Semi-supervised dissimilarity representation

Let T and S be the training and test sets. Let R be the representation set composed of k prototypes, $R = \{r_1, r_2, \ldots, r_k\}$. An object x belonging to T or S is

[*] Secondary affiliation with the Carinthian Tech Research AG, Villach, Austria.

978-4-9906441-1-6 ©2012 IAPR

represented as: $d_x = [d(x, r_1), d(x, r_2), \ldots, d(x, r_k)]$ in which $d(x, r_i), i = 1 \ldots k$, is the distance between x and the prototype r_i. $d(x, r_i)$ can also be seen as features of x in the constructed dissimilarity space. Based on the constructed dissimilarity space, standard machine learning techniques can be applied to perform the classification.

In the supervised setting, the representation set is traditionally selected from the training set ($R \subset T$). In our semi-supervised setting, the representation set also includes objects from the test set ($R \subset \{T \cup S\}$). The training and test sets in both configurations are the same, only the object representation changes. Note that we use test data for object representation and disregard their labels during training.

Our semi-supervised method for the dissimilarity representation can be categorized as the "Change of Representation" approach [2] which aims at enhancing the data representation using unlabeled data. We enhance the data representation by enlarging the representation set to capture different aspects of the data and thus providing more discriminative information for the classification task under consideration.

We investigate the performance of the supervised and semi-supervised settings in two scenarios:

- All of the available objects are used to build the representation set

$$R := \begin{cases} T & \text{in the supervised setting} \\ T \cup S & \text{in the semi-supervised setting} \end{cases}$$

- A set of informative prototypes is selected, e.g. by using a feature selection technique, from the available prototypes/objects. In this scenario, we examine whether feature selection benefits from the enlargement of its search space to unlabeled data.

3. Experiments

We use two base classifiers: the linear Support Vector Machine (LSVM) with a default tradeoff parameter value (C = 1) and the k-NN classifier with $k = 1$ (1-NN). We divide each data set into training and test sets of various sizes. At each time, the training set is selected randomly based on the data distribution. We repeat the experiments 150 times and average the classification results.

3.1 Data sets

We have selected standard data sets from two and ten class classification problems. The distance between object is measured in various ways. In this paper, we present four data sets: Polygon, 38-haus, 38-eucl, and Zongker as described in Table 1. The polygon data set [5] consists of two classes of randomly generated polygons: 2000 convex quadrilaterals and 2000 irregular heptagons. The polygons are first scaled and then their similarity is computed as the modified Hausdorff distances between their vertices. The Zongker data set comes from the NIST digits, originally given as 128 × 128 binary images [6]. The data set is composed of 10 classes, each class consists of 200 samples. The deformable template matching defined by [4] is used as the similarity measure. The 38-haus and 38-eucl data sets also come from the NIST digits but only consider the images of digits '3' and '8'. Each digit class consists of 1000 samples. The similarity measures used in the two data sets are the Hausdorff and Euclidean distances, respectively.

Table 1. Data sets used in experiments

Data	Distance Measure	#Samples
Polygon	Modified Hausdorff	2×2000
38-haus	Hausdorff	2×1000
38-eucl	Euclidean	2×1000
Zongker	Template-Matching	10×200

3.2 Using all available prototypes

The results with respect to different training sizes using the supervised **(SU)** and semi-supervised **(SE)** settings on the four data sets are shown in Figure 1. Averaged error rate over 150 repetitions (vertical axis) is plotted against the training set size (horizontal axis) by dashed line for the supervised setting and by solid line for the semi-supervised setting. Red and black display the results for LSVM and 1-NN, respectively.

The plots show that when the LSVM is used, the semi-supervised setting often outperforms the supervised setting. The semi-supervised setting performs better than the supervised setting for three data sets Polygon, 38-NIST using Hausdorff distance (38-haus), and Zongker but worse for the 38-NIST data set using the Euclidean distance (38-eulc). The difference in performance between the two settings manifests clearly if the training set size is small. For example, the semi-supervised setting yields a decrease of 14% error rate compared with the supervised setting if the training set size is 30 for the Zongker data set, and a decrease of 8% error rate if the training set size is 10 for the Polygon data set. This verifies our statement that in the case of limited training data, objects are better described by including more prototypes from the test set into the representation set. When the training set is sufficient large, the two settings have similar performance since in that case the representation set in the supervised setting is large enough to describe the data.

When the 1-NN classifier is used, the semi-supervised setting works slightly worse than the supervised setting for the 38-NIST data set using the Hausdorff or the Euclidean distance. This behavior may be due to the rather high-dimensionality of the dissimilarity space with which the 1-NN has difficulty in dealing [1]. In such a situation, employing a feature selection step might improve the performance of the classifier. This is indeed demonstrated in Section 3.3.

The advantage of the semi-supervised setting is further demonstrated by varying the sizes of the training set and the representation set. Figure 2 presents the results for the Polygon data using the LSVM. The horizontal axis shows the size of the representation set; colors represent the size of the training set. Note that if the size of the representation set is larger than or equal to that of the training set, then the representation set first includes all samples from the training set and the rest is randomly selected from the test set. If the size of the representation set is smaller than that of the training set, then samples of the representation set are randomly selected from the training set. Thus, the supervised setting is equivalent to the case in which the size of the representation set is equal to the size of the training set. As shown in Figure 2, for a fixed training set size, increasing the representation set size leads to the improvement in the classification result. The results become stable when the representation set size is large, i.e. larger than 400 or 10% of the whole data set in this case. It is worth noting that it is unnecessary to select all samples for representation; for this Polygon data set, just 10% of the data is enough to achieve good classification result.

3.3 Selecting prototypes from the available set

We use the prototype selection, which employs linear programming, for dissimilarity-based classifiers as presented in [5]. This method recasts the prototype selection as a classification problem which aims at determining a two-class sparse linear classifier. The prototypes associated with the non-zero weights of the classifier are then selected for the "final" representation set.

Figure 3 shows the classification results for the 38-haus data set where we first employ prototype selection and then 1-NN classifier. The average error rate (vertical axis) is plotted against the size of the training set (horizontal axis) as in Figure 1. Results with and without prototype selection are displayed in red and black.

The semi-supervised setting without prototype selection implies that all available prototypes (objects) are used for the representation. As shown in the figure, prototype selection leads to better classification result. In the supervised setting without prototype selection, just the training set is used for the representation. On the contrary to the semi-supervised setting, the classification result with prototype selection is far worse than without prototype selection if the training set size is less than 100. It is because the small representation set makes it difficult for the prototype selection method to select "good" prototypes for the "final" representation set. It should be noted that if the training set size is larger than 17, the 1-NN classification with prototype selection under the semi-supervised setting performs best.

4. Discussion and Conclusions

This paper shows a study of dissimilarity-based classification where the representation set is enlarged by including samples from the test set. As a result, the representation set provides a richer description for the objects of interest and helps to significantly improve the classification performance, especially for small sample size problem. Therefore, the semi-supervised setting is helpful for situations in which the availability of labelled data is limited.

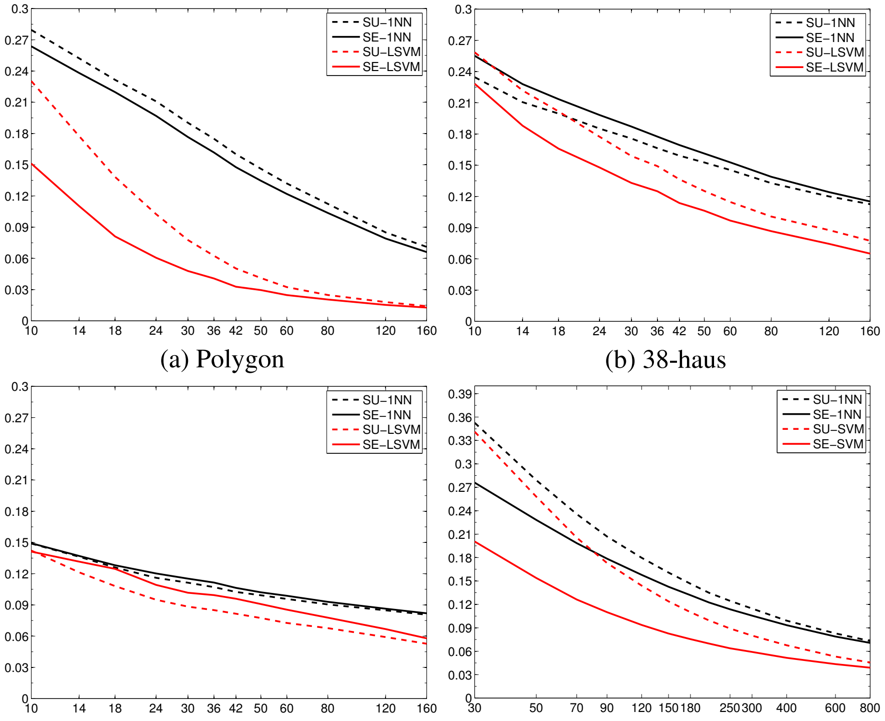
If all available objects are included into the representation set, the computation of the distance matrix might, however, become expensive, or in other ways prohibit the direct use of the dissimilarity method. Nevertheless, it is shown in the experiments that in many cases using a subset of the data for the representation provides as good classification result as using all data (cf. Figure 2). Prototype selection methods might be used to select such a representation set (Figure 3) and, in many cases, still improve upon a representation set merely derived from the training data.

We, however, note that the semi-supervised setting does not always perform better than the supervised setting, e.g. for the 38-eucl data set. It is of future interest to further investigate in which situations the semi-supervised setting does help dissimilarity-based classifiers improve upon the standard supervised setting.

References

[1] C. Aggarwal, A. Hinneburg, and D. Keim. On the surprising behavior of distance metrics in high dimensional space. *Database Theory*, pages 420–434, 2001.

[2] O. Chapelle, B. Scholkopf, and A. Zien, editors. *Semi-supervised Learning*. MIT Press, 2006.



(a) Polygon

(b) 38-haus

(c) 38-eucl

(d) Zongker

Figure 1. Classification results for the four data sets using the LSVM and 1-NN classifiers. SU and SE stand for supervised and semi-supervised settings, respectively. The averaged error rate over 150 repetitions (vertical axis) is plotted against the training set size (horizontal axis).

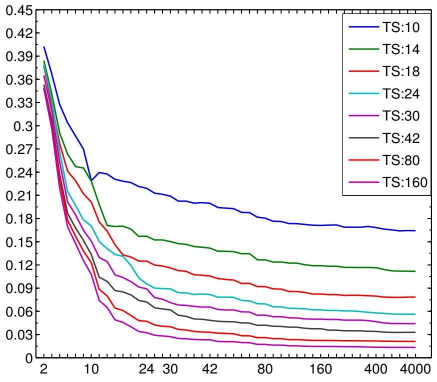


Figure 2. Results on the Polygon data with varying representation set and training set sizes.

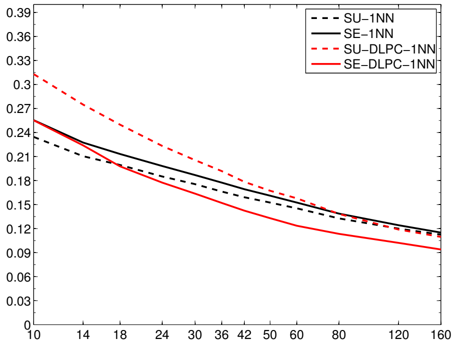


Figure 3. Results for the 38-haus data using 1-NN classifier with and without prototype selection.

[3] R. Duin and E. Pekalska. The dissimilarity representation for structural pattern recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7042, pages 1–24, 2011.

[4] A. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *Pattern Analysis and Machine Intelligence*, 1997.

[5] E. Pekalska, R. Duin, and P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, 2006.

[6] C. Wilson and M. Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, 1992.