Expert Systems with Applications 40 (2013) 7069–7079

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Semi-supervised object recognition based on Connected Image Transformations

Steven Van Vaerenbergh [a,*], Ignacio Santamaría [a], Paolo Emilio Barbano [b]

[a] Dept. of Communications Engineering, University of Cantabria, Plaza de la Ciencia s/n, Santander, 39005, Spain
[b] Dept. of Mathematics, Yale University, PO Box 208283, New Haven, CT 06520-8283, USA

ARTICLE INFO

Keywords:
Semi-supervised classification
Object recognition
Connectivity
Deformation models
Low-density separation

ABSTRACT

We present a novel semi-supervised classifier model based on paths between unlabeled and labeled data through a sequence of local pattern transformations. A reliable measure of path-length is proposed that combines a local dissimilarity measure between consecutive patters along a path with a global, connectivity-based metric. We apply this model to problems of object recognition, for which we propose a practical classification algorithm based on sequences of "Connected Image Transformations" (CIT). Experimental results on four popular image benchmarks demonstrate how the proposed CIT classifier outperforms state-of-the-art semi-supervised techniques. The results are particularly significant when only a very small number of labeled patterns is available: the proposed algorithm obtains a generalization error of 4.57% on the MNIST data set trained on 2000 randomly chosen patterns with only 10 labeled patterns per digit class.

1. Introduction

In many object recognition problems, obtaining labeled data is a time-consuming and expensive task, whereas large unlabeled data sets are usually available. This is particularly true in problems involving high-dimensional data, such as handwritten digit recognition, text categorization (Joachims, 1998), protein classification (Weston, Leslie, Zhou, Elisseeff, & Noble, 2004) or hyper-spectral data classification (Rajan, Ghosh, & Crawford, 2008). In such scenarios it is desirable to develop semi-supervised learning techniques, as these allow to exploit the available unlabeled data concurrently with the labeled training data. We focus on recognition problems in which many instances of each object are available for training, and each instance differs only slightly from another instance of the same object. This is typically the case in handwritten digit and face recognition systems, but it occurs more generally in a wide variety of image recognition problems where series of spatially or temporally related patterns are available. In this case, the available instances of an object usually relate to each other by transformations such as rotations, scalings and small nonlinear axis deformations.

A large number of semi-supervised learning techniques have been proposed in the last years, for instance (Belkin, Niyogi, & Sindhwani, 2006; Cohen, Cozman, Sebe, Cirelo, & Huang, 2004; Fischer, Roth, & Buhmann, 2004; Jaakkola & Szummer, 2002; Szlam, Maggioni, & Coifman, 2008; Wang & Zhang, 2007; Zhu, 2005; Zhu, Ghahramani, & Lafferty, 2003; Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004). The success of these techniques relies mainly on two key assumptions: (i) the data lie on a manifold of much lower dimensionality than the data dimension itself (manifold assumption) (Belkin et al., 2006); and (ii) data points belonging to the same high-density region are likely to belong to the same class (cluster assumption) (Fischer et al., 2004). Both assumptions can be interpreted in terms of data similarity and distances. In this sense, the manifold assumption states that *local* variations in the data should only involve variations of a small number of parameters. This property is illustrated in Fig. 1, which shows a number of handwritten instances of the number 3: although the data dimensionality is high, most local variations can be described by few parameters, such as line thickness, skew and rotation. Therefore, the manifold assumption leads naturally to the concept of a *local* distance between patterns. Several algorithms exploit the manifold assumption, e.g. (Belkin & Niyogi, 2002), by estimating the marginal distribution underlying the data and training a classifier on the manifold itself.

The cluster assumption states that two data points should belong to the same class if they can be connected by a path that lies exclusively in a region of high density. This assumption, which was exploited for instance in Fischer et al. (2004) and Chapelle and Zien (2005), allows to define a *global* distance measure between patterns that lie further apart. Specifically, the global distance

* Corresponding author. Tel.: +34 942200919x802; fax: +34 942201488.
E-mail addresses: steven@gtas.dicom.unican.es (S. Van Vaerenbergh), ignacio@gtas.dicom.unican.es (I. Santamaría), paoloemilio.barbano@yale.edu (P.E. Barbano).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.029

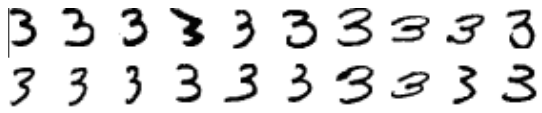

Fig. 1. Handwritten instances of the number 3, from the MNIST data set.

between two points is measured as the length of the path between them, in which each connection is measured as a local dissimilarity between two intermediate patterns. Therefore, while the manifold assumption refers to a local dissimilarity, the cluster assumption refers to a global distance.

The proposed semi-supervised method uses small pattern transformations as the local dissimilarity measure. They are accumulated along a path using a connectivity distance to obtain a robust and reliable global distance, and a simple nearest-neighbor technique is finally used for classification. Despite of its fairly simple formulation, the new algorithm outperforms state-of-the-art semi-supervised classification algorithms when tested on standard benchmark image data sets. Some preliminary results of the proposed method appeared in Van Vaerenbergh, Santamaría, and Barbano (2011). Here, we extend the experimental study of the algorithm, and we formulate the out-of-sample classification procedure. The algorithm has quadratic time and memory complexity in terms of the number of training points, which becomes impractical if large data sets are used. In order to reduce the out-of-sample classification cost, we also propose a prototype-based approximation procedure.

The paper is organized as follows: in Section 2, we provide some related literature and previous work. In Section 3, we review the semi-supervised classification setting and state the main assumptions on which the proposed method is based. Section 4 introduces the local and global dissimilarity measures, which form the basis of the proposed semi-supervised classifier, and it describes the proposed technique. An out-of-sample extension is discussed in Section 5, including a strategy for dealing with large-scale data sets. Section 6 illustrates the obtained performance in comparison to other state-of-the-art techniques on four typical databases. Finally, we summarize the main conclusions of this work in Section 7.

2. Previous work on semi-supervised classification

Much of the recent effort in semi-supervised learning has been centered around the problem of finding a reliable method to infer a global distance measure from local dissimilarities (Belkin & Niyogi, 2002; Belkin et al., 2006; Jaakkola & Szummer, 2002; Szlam et al., 2008; Zhu et al., 2003; Zhou et al., 2004), which is also the main problem addressed in the present contribution. Most of these techniques start by constructing an undirected weighted graph (or, equivalently, an affinity matrix) on the labeled and unlabeled data points, where the edge weights measure the pairwise dissimilarities. Then, they apply different approaches to design a global classifying function with desirable properties (e.g., smoothness, robustness, etc.). For instance, Jaakkola and Szummer (2002) and Szlam et al. (2008) use a probabilistic approach in which the graph weights (local dissimilarities) are viewed as transition probabilities and the global dissimilarities are established through a random walk or a diffusion process on the graph, respectively. The local dissimilarity metric in Jaakkola and Szummer (2002) and Szlam et al. (2008), however, is computed by the standard Gaussian kernel, which makes the estimate of the shortest path length more sensitive to noise. Instead of considering just the shortest path, these algorithms integrate the volume of all paths between two data points, hence effectively de-noising the global metric.

Closely related approaches that eliminate the dependency of Jaakkola and Szummer (2002) and Szlam et al. (2008) with respect to the diffusion time are the harmonic Gaussian field classifier described in Zhu et al. (2003) and the consistency method in Zhou et al. (2004). These methods estimate a global metric on the weighted graph (i.e., the semi-supervised classifier) by repeatedly applying the Laplacian matrix (or some of its normalized versions) over a matrix of labels which is consistent with the training data. Over iterations, label information is propagated through the graph and, after reaching a stable state, the unlabeled patterns are assigned to the classes from which they have received more information. Again, these methods use the conventional Gaussian kernel as the local similarity function for computing the affinity matrix. The smoothness constraint imposed by the Laplacian is in this case responsible for de-noising the global metric.

In Belkin et al. (2006) proposed a framework that exploits the geometry of the underlying marginal distribution, which can be estimated from unlabeled data, to regularize the data manifold. This principle was used to design a semi-supervised classifier, denoted as the Laplacian Support Vector Machine (LapSVM). The resulting classifier has the interesting property of providing a natural out-of-sample extension. In order to lower the cubic training complexity of LapSVM, a training algorithm in the primal was recently proposed in Melacci and Belkin (2011).

While many other graph-based approaches for semi-supervised classification have been proposed over the past years, all of them use for the local dissimilarities a function of the Euclidean distance with exponential decay, typically the Gaussian kernel, regardless of the particular application considered. Their emphasis is on how a suitable global metric or function for semi-supervised learning should be estimated from a graph, and to this end they proposed quite sophisticated methods. Departing from that trend, in this work we demonstrate that better results can be obtained by translating most of the complexity to the computation of the local dissimilarity measure. This metric should be problem-dependent to better characterize the data manifold structure at a local scale. In doing so, we can simplify the global metric as a shortest path which can be implemented using Dijkstra's algorithm, as we will show below. This conceptually simple procedure provides very good results in different scenarios.

3. Problem formulation and assumptions

We consider a multi-class classification problem with N classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. In a semi-supervised classification setting, we are given a training data set consisting of $n = l + u$ patterns, $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1} \ldots, \mathbf{x}_n\} = \mathcal{X}_l \cup \mathcal{X}_u$, represented as Euclidean vectors of dimension dim. The first l patterns in this set correspond to labeled data, with class labels $\{y_1, \ldots, y_l\}$, while the remaining u patterns constitute the available unlabeled data. We assume that all input patterns $\mathbf{x}_i$ have been drawn independently and identically distributed (i.i.d.) from some unknown marginal data distribution $P(\mathbf{x})$. This is the conventional setting assumed in most semi-supervised classification techniques described in the literature (Chapelle et al., 2006; Cohen et al., 2004). Furthermore, in this paper we are interested in semi-supervised classification problems where l is a very small fraction of the total number of available patterns, n.

We now make the following assumptions:

Assumption 1. For each pattern $\mathbf{x}_i$ in the data set there exist close patterns $\mathbf{x}_j$ of the same class ($y_i = y_j$) that can be obtained by small transformations of the given pattern. The norm of these transformations is measured by some intrinsic dissimilarity measure.

Assumption 2. For any two patterns $\mathbf{x}_i$ and $\mathbf{x}_j$ that belong to the same class ($y_i = y_j$), there exists a sequence of k transformations