Pattern Recognition Letters 37 (2014) 24–31

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Semi-supervised linear discriminant analysis through moment-constraint parameter estimation

Marco Loog *

Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
The Image Group, University of Copenhagen, Universitetsparken 5, DK-2100, Copenhagen Ø, Denmark

ARTICLE INFO

Article history:
Available online 18 March 2013
Communicated by G. Sanniti di Baja

Keywords:
Linear discriminant analysis
Semi-supervised learning
Moment constraints
Affine invariant
Classification

ABSTRACT

A semi-supervised version of classical linear discriminant analysis is presented. As opposed to most current approaches to semi-supervised learning, no additional extrinsic assumptions are made to tie information coming from labeled and unlabeled data together. Our approach exploits the fact that the parameters that are to be estimated fulfill particular relations, intrinsic to the classifier, that link label-dependent with label-independent quantities. In this way, the latter type of parameters, which can be estimated based on unlabeled data, impose constraints on the former and lead to a reduction in variability of the label dependent estimates. As a result, the performance of our semi-supervised linear discriminant is typically expected to improve over that of its regular supervised match. Possibly more important, our semi-supervised linear discriminant analysis does not show the severe deteriorations other approaches frequently display with increasing numbers of unlabeled data. This work recapitulates, corrects, extends, and revises our previous work that has been published as part of the First IAPR TC3 Workshop on Partially Supervised Learning. The main novelty it provides over our earlier work is an affine invariant approach to semi-supervised learning befitting linear discriminant analysis. Besides, more elaborate and convincing experimental evidence of the potential of our general approach is provided. We essentially believe that the general principle of intrinsic constraints is of interest as such and may inspire other novel semi-supervised methods.

1. Introduction

Supervised learning aims to learn from examples: given a limited number of instances of a particular input–output relation, its goal is to generalize this relationship to new and unseen data in order to enable the prediction of the associated output given new input. Specifically, supervised classification seeks to infer an unknown feature vector-class label relation from a finite number of feature vectors and their associated, desired class labels. Now, an elementary question is whether and, if so, how additional unlabeled data can significantly improve the training of such classifier. This is what constitutes the problem of semi-supervised learning (Chapelle et al., 2006; Zhu and Goldberg, 2009).

The expectation is that semi-supervised learning can indeed bring considerable improvements to many research and application areas in which classification problems play a key role by simply exploiting the often enormous amounts of unlabeled data available (think image analysis, computer vision, natural language processing, medical diagnostics, but also the social and environmental sciences and various metrics). The matter of the fact, however, is that current semi-supervised methods have not been widely accepted outside of the realms of computer science. Part of the reason for this may be that current methods offer no performance guarantees (Ben-David et al., 2008; Singh et al., 2009) and often deteriorate when confronted with large amounts of unlabeled samples (Cohen et al., 2004; Cozman and Cohen, 2006; Mann and McCallum, 2010; Nigam et al., 1998).

Earlier, we identified as main reason for the frequent failure of semi-supervised learning the fact that current semi-supervised approaches typically rely on assumptions extraneous to the classifier being considered (Loog, 2010, 2012). Indeed, the main current approaches to semi-supervised learning stress the need for extrinsic assumptions such as the cluster assumption: points from the same class cluster, the smoothness assumption: neighboring point have the same label, the assumption of low density separation: the decision boundary is located in low density areas, and the like (Chapelle et al., 2006; Zhu and Goldberg, 2009). Given a particular assumption holds, one is able to extract relevant information not only from the labeled, but especially from the unlabeled examples. While it is undeniably true that having more precise knowledge on the distribution of data could, or even should, help in training a better classifier, in many real-world settings it may be questionable if one can

* Address: Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands. Tel.: +31 15 27 89395.
E-mail address: m.loog@tudelft.nl
URL: http://prlab.tudelft.nl

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.03.004

at all check if such conditions are indeed met. Moreover, as soon as these additional model assumptions do not fit the data, there obviously is the real risk that adding unlabeled data actually leads to a severe deterioration of classification performance (Cohen et al., 2004; Cozman and Cohen, 2006; Loog, 2010, 2012; Nigam et al., 1998). Note that this is in contrast with the supervised setting, where most classifiers, generative or not, are capable of handling mismatched data assumptions rather well, in the sense that adding more training data generally improves the performance of the classifier (but cf. Loog and Duin, 2012).

The current work devises a specific semi-supervised scheme tailored to classical linear discriminant analysis (LDA, Hastie et al., 2001; Ripley, 1996), which is sometimes referred to as a normal-based linear discriminant function (see McLachlan, 1992). Regarding LDA, we would like to stress that it still is a widely employed classifier and therefore, also from a practical perspective, the current investigation is of interest and could have consequences beyond the mere academical. Also, like any other classifier, LDA has its validity and cannot be put aside as being outdated or not state-of-the-art. In this respect, we would also like to refer the reader to insightful contributions such at the ones by Hand (2006) and Efron (2001) (see also Duin et al., 2010).

Now, the underlying, more general idea presented in this paper is that the class-specific parameters to be estimated in the learning phase are related to each other and, more importantly, to certain label-independent statistics. These relations can be seen as intrinsic—rather than additional, extrinsic—constraints between particular estimates coming from labeled data and those derived from unlabeled instances. Enforcing these constraints during semi-supervised learning yields label-dependent estimates that are in a sense closer to the true parameter values, which, in turn, often lead to reduced classification errors.

Though the focus in this work is specifically on LDA, we do believe that the general, underlying principle of searching for intrinsically motivated semi-supervised learning is of interest in its own respect and we hope that it this work will inspire further research in this direction.

On the conceptual side, this work continuous in the spirit of the earlier research reported on in Loog (2010, 2012). Methodologically, the paper presents a revised version of the latter contribution and extends and corrects parts of the method proposed there. In particular, it dwells on an important shortcoming of the technique from Loog (2012), which is the lack of invariance (or, if one prefers, covariance) of the parameters of semi-supervised LDA under nonsingular affine transformations of the feature space. Such affine transformations do not only comprise translation, rotation, reflection, and isotropic scaling but also anisotropic scaling, and shearing. Classification methods have this invariance property if, for any two nonsingular affine transformation A and B, the classifiers C_A and C_B trained on data transformed by A and B, respectively, deliver the same classification outcomes on a similarly transformed test object x, i.e., $C_A(A(x)) = C_B(B(x))$. Note that there are many classifiers for which this noticeable property does not hold, e.g. neither k-nearest neighbors, nor Parzen classifiers, nearest mean classifiers, or support vector machines are invariant to anisotropic scaling and shearing transforms. The regular counterpart to semi-supervised LDA, however, does enjoy this invariance property (Fukunaga, 1990; McLachlan, 1992). It therefore seems nothing more than reasonable to retain the same invariance for the semi-supervised case. This paper demonstrates how such elementary characteristic can also be enforced in semi-supervised learning.

1.1. Outline

Following the next section, which presents an overview of further related work, Section 3 briefly recapitulates some relevant details of the approach presented in Loog (2010) for semi-supervised nearest mean classification. The main focus in that section will, however, be on semi-supervised LDA as presented in Loog (2012) and its novel variation that satisfies the earlier sketched invariance property. Section 4 provides experimental results on various real-world and benchmark data sets in which our constrained approach is mainly compared to regular LDA and so-called self-learned LDA (the latter of which is briefly explained in the next section as well). Additional comparisons are made with logistic regression, nearest neighbor classification, an entropy regularization method, and transductive SVM. Subsequently, Section 5 completes the paper, providing a discussion and conclusions.

2. Additional related works

There are few works that focus on semi-supervised LDA (see Efron (2001)'s rule 1). Most relevant contributions come from statistics and have been published mainly by the end of the 1960s and halfway the 1970s. Hartley and Rao (1968) suggests to maximize the likelihood over all permutations of possible labelings of unlabeled objects. A computationally more feasible approach has been proposed by McLachlan (1975, 1977), which follow an iterative procedure. Firstly, the linear discriminant is trained on the labeled data only and used to label all unlabeled instances. Using the now-labeled data, the classifier is retrained and employed to relabel the initially unlabeled data. This process of relabeling originally unlabeled data is repeated until none of the samples changes label.

The above approach to semi-supervised learning is basically a form of so-called self-training or self-learning, which has been presented in different guises and at different levels of complication (see, for instance, Basu et al., 2002; McLachlan, 1975; McLachlan and Ganesalingam, 1982; Nigam et al., 1998; Titterington, 1976; Vittaut et al., 2002; Yarowsky, 1995; Zhou and Li, 2010). This iterative method also relates directly to the well-known approach to semi-supervised learning based on expectation maximization (see Nigam et al. (1998) and the discussion papers related to Dempster et al. (1977)). The similarity between self-learning and expectation maximization (in some cases equivalence even) has been noted in various papers, e.g. by Abney (2004), Basu et al. (2002), and it is to no surprise that such approaches suffer from the same drawback: as soon as the underlying model assumptions do not fit the data, there is the real risk that adding too much unlabeled data leads to a substantial decrease of classification performance (Cohen et al., 2004; Cozman and Cohen, 2006; Nigam et al., 1998).

An approach seemingly different from self-learning is, among others, known as label propagation. It relies on the smoothness assumption, assuming that data points close to each other tend to belong to the same class. Various versions of this idea have been studied, most of which are related to graph-based techniques, manifold learning, or spectral clustering methods (Bengio et al., 2006; Szummer and Jaakkola, 2002; Zhu and Ghahramani, 2002). The propagation of label information through such graph structure can also be thought of as a particular instantiation of the iterative expectation maximization or self-learning methods. A more explicit connection between self-learning and graph-based propagation methods can be found in Culp and Michailidis (2008).

We finally remark that there are also semi-supervised approaches to LDA as a dimensionality reduction technique. As we consider LDA as a classifier, we do not discuss these approach in any detail. As it comes in some sense close to our work, the single paper we do like to mention is by Fan et al. (2009). The work notes that the Fisher criterion, which typically employs the between-class and within-class covariance matrices, can also be expressed in such a way that the total covariance matrix replaces one of