Signal Processing 90 (2010) 236–248

Contents lists available at ScienceDirect

Signal Processing

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

Large margin nearest local mean classifier

Jing Chai [a,*], Hongwei Liu [a], Bo Chen [b], Zheng Bao [a]

[a] National Laboratory of Radar Signal Processing, Xidian University, Xi'an, Shanxi, China
[b] Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA

ARTICLE INFO

Article history:
Received 11 November 2008
Received in revised form
8 June 2009
Accepted 10 June 2009
Available online 17 June 2009

Keywords:
Large margin
Local mean vector
Linear transformation
Eigenvalue decomposition
Dimension
Semidefinite programming

ABSTRACT

Distance metric learning and classifier design are two highly challenging tasks in the machine learning community. In this paper we propose a new large margin nearest local mean (LMNLM) scheme to consider them jointly, which aims at improving the separability between local parts of different classes. We adopt 'local mean vector' as the basic classification model, and then through linear transformation, large margins between heterogeneous local parts are introduced. Moreover, by eigenvalue decomposition, we may also reduce data's dimensions. LMNLM can be formulated as a semidefinite programming (SDP) problem, so it is assured to converge globally. Experimental results show that LMNLM is a promising algorithm due to its leading to high classification accuracies and low dimensions.

1. Introduction

Distance metric learning is of fundamental importance to many machine learning tasks, such as supervised classification, unsupervised clustering, semi-supervised learning, image retrieval, etc. [1]. A good distance metric may provide us with some enlightenment on underlying data structures, and thus an improved performance is expected to be obtained with the learned metric. In this paper we concentrate on supervised classification problems, so if a good metric can be learned effectively, higher classification accuracies will be achieved.

Another important aspect in machine learning is how to design classifiers. Similar to distance metric learning, a good classifier should also be data-dependent, i.e., some implicit data structures should be revealed by the designed classifier. Due to the common requirement of metric learning and classifier design, the two seemingly different problems could be considered jointly. Moreover, although there are some well-established classification approaches such as support vector machines (SVMs) [2,3] and neural networks [4], most of these approaches do not consider the problem of learning a proper distance metric that is suitable for the corresponding classification model. Motivated by the above analysis, in this paper, we combine distance metric learning and classifier design as a whole, and this combination is one of central guidelines for our work.

In view of minimizing the structural risk, a special kind of 'large margin' classifier, support vector machines [2,3], achieved general success in the last decade. Recently, some metric learning and/or classifier design works have adopted ideas similar to 'large margin' of SVMs. Xing et al. [5] applied the side information for Mahalanobis distance metric learning and aimed at minimizing the sum of squared distances between similarly labeled pairs and simultaneously maintaining a lower bound on the sum of squared distances between dissimilarly labeled ones. Although can be solved with global optimal solutions, the algorithm of [5] does not involve any slack variables and may incur overfittings. Bachrach et al. [6] introduced the 'margin' into feature selection tasks. The algorithm of [6] cannot be described as a convex optimizing problem and usually converges with local optimal solutions. The

* Corresponding author. Tel.: +86 29 88206441; fax: +86 29 88201448.
E-mail address: jingchai@yahoo.cn (J. Chai).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.06.015

work in [7–10] concentrated on different 'marginal' metric learning algorithms for k-NN classifiers. They employ the k-NN classifier as classification tools, which utilize information from k nearest neighbors and may neglect the discriminative information derived from other useful inputs. Veenman et al. [11] pioneered the combination of feature weighting and the nearest mean (NM) classifier. One obvious limitation of this feature weighting algorithm is that it is only suitable for binary classification problems. Jenssen et al. [12] designed an NM classifier which is based on nonlinear transformations. Peltonen et al. [13] introduced a probabilistic model to generalize linear discriminant analysis (LDA) [14,15] for selecting informative or relevant components. The realization of both [12,13] needs to evaluate the class probability density function with Parzen window, and consequently, the number of training examples used for the evaluation should be large enough to assure a high evaluation accuracy. Therefore, the algorithms of [12,13] may be unsuitable for applications when the small sample size (SSS) problem occurs. Yang et al. [16] proposed a distance metric learning algorithm that optimizes local compactness and separability by a probabilistic framework. The algorithm of [16] is not convex and often leads to local optimal solutions.

Our work in this paper is also motivated by the 'large margin' enlightenment. In particular, here we propose a large margin nearest local mean (abbreviated as 'LMNLM') classifier which introduces large margins between local parts of different target classes and expects to improve the class separability. To reach this goal, we replace the old Euclidean distance metric by a new Mahalanobis one through linear transformations. Fortunately, we can express our algorithm as a semidefinite programming (SDP) problem, which is one special kind of convex optimizing problems. Due to its convex nature, the proposed LMNLM algorithm can avoid the trouble of local minima since it is assured to converge globally.

Dimensionality reduction is also quite important in the machine learning community. With the help of dimensionality reduction, three advantages are expected to be obtained in pattern recognition tasks. Firstly, we can save the memory space and reduce the system complexity. Secondly, those useless and harmful noisy components can be removed and improved classification accuracies can be obtained. Thirdly, it makes us capable to deal with data set with few samples and high dimensions and thus weaken the disadvantages caused by the so-called curse-of-dimensionality problem. In this paper, we also consider the dimensionality reduction problem, and how the proposed algorithm leads to lower-dimensional solutions is also introduced.

The rest of this paper is organized as follows. In Section 2, we discuss how the LMNLM classifier is designed and how the dimensions are reduced in detail. Some related work and their relations to our work are discussed in Section 3. In Section 4, we conduct classification experiments, respectively, on synthetic data set, benchmark data sets, and radar high-resolution range profiles (HRRP) data set, and then an analysis of these results follows. Finally, we give concluding remarks and suggestions for future work in Section 5.

2. The LMNLM scheme

Suppose that we are given a set of data points $\{x_k|x_k \in R^D\}$ with $k \in \{1,2,\ldots,n\}$, and that $y_k \in \{1,2,\ldots,c\}$ denotes the class label for x_k. $\{\mu_1,\ldots,\mu_c\}$ are the corresponding mean vectors for samples in c classes. Given a testing point z, $d^2(z,\mu) = \|z-\mu\|_2^2$ denotes the squared Euclidean distance between z and μ. Based on the above notations, the nearest mean (NM) classifier can be defined as follows: we decide that the testing point z belongs to class g, if for $\forall h \in \{1,2,\ldots,c\}$, $d^2(z,\mu_g) \le d^2(z,\mu_h)$ is always satisfied.

The NM classifier is simply modeled with class mean vectors. Under the assumption of statistically independent features with equal covariance, i.e. $\sum_1 = \sum_2 = \cdots = \sum_c = \sigma^2 I$, Euclidian distance is optimal with respect to Bayesian criterion. However, the above assumptions are too strict to satisfy in most real-world applications. Firstly, the assumption of equal covariance may be unsatisfied, i.e., $\sum_1 \neq \sum_2 \neq \cdots \neq \sum_c$. Secondly, the NM classifier is only suitable for single modal distributed data. However, in practice, many data are multimodal distributed. Thirdly, the assumption that different features have the same variance may be unsatisfied, i.e., different diagonal components of $\sum_i$ may be unequal to each other. Fourthly, features may be statistically dependent, i.e., $\sum_i$ may be not a diagonal matrix but with non-zero non-diagonal elements, and in this situation, cross-correlations among different features should be considered as well. For the above third and fourth case, adopting the Mahalanobis distance will be a better choice than adopting the Euclidean one, because the former can avoid the limitation that different features are statistically independent with equal variance. To treat with the limitation in the above first case, we can either employ different covariance matrix for each class, e.g., utilizing the multivariate Gaussian classifier, or change the 'mean vector' classification model into another one that does not require the covariance matrix for each class to be the same. However, for the limitation in the above second case, due to it is caused by the intrinsic 'mean vector' model of the NM classifier, we can only employ other classification models to avoid this limitation.

According to the above analysis, an alternative model, the 'local mean vector' one, is proposed to replace the old 'mean vector' one. With this new model, the limitations in the first and second cases are expected to be released and even avoided. Next we give the definition of our new model.

Definition 1. The local mean vector.

Suppose that the affinity matrix $G \in R^{n\times n}$ is calculated as $G_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ for $i \neq j$ and $G_{ii} = 0$, where σ is the kernel width parameter. Then x_k's local mean vector for samples in class l is defined as

$$m_{kl} = \frac{\sum_{y_i=l} G_{ki} x_i}{\sum_{y_i=l} G_{ki}} \tag{1}$$