Neural Networks 22 (2009) 977–987

Contents lists available at ScienceDirect

Neural Networks

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

A novel kernel-based maximum a posteriori classification method

Zenglin Xu [a], Kaizhu Huang [b], Jianke Zhu [a], Irwin King [a,*], Michael R. Lyu [a]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

[b] Department of Engineering Mathematics, University Of Bristol, Bristol, BS8 1TR, United Kingdom

ARTICLE INFO

Article history:
Received 19 March 2008
Revised and accepted 20 November 2008

Keywords:
Kernel methods
Maximum a posteriori
Discriminant analysis

ABSTRACT

Kernel methods have been widely used in pattern recognition. Many kernel classifiers such as Support Vector Machines (SVM) assume that data can be separated by a hyperplane in the kernel-induced feature space. These methods do not consider the data distribution and are difficult to output the probabilities or confidences for classification. This paper proposes a novel Kernel-based Maximum A Posteriori (KMAP) classification method, which makes a Gaussian distribution assumption instead of a linear separable assumption in the feature space. Robust methods are further proposed to estimate the probability densities, and the kernel trick is utilized to calculate our model. The model is theoretically and empirically important in the sense that: (1) it presents a more generalized classification model than other kernel-based algorithms, e.g., Kernel Fisher Discriminant Analysis (KFDA); (2) it can output probability or confidence for classification, therefore providing potential for reasoning under uncertainty; and (3) multi-way classification is as straightforward as binary classification in this model, because only probability calculation is involved and no one-against-one or one-against-others voting is needed. Moreover, we conduct an extensive experimental comparison with state-of-the-art classification methods, such as SVM and KFDA, on both eight UCI benchmark data sets and three face data sets. The results demonstrate that KMAP achieves very promising performance against other models.

1. Introduction

Kernel methods play an important role in machine learning and pattern recognition (Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004). They have achieved success in almost all traditional tasks of machine learning, i.e., supervised learning (Mika, Ratsch, Weston, Scholkopf, & Muller, 1999; Vapnik, 1998), unsupervised learning (Schölkopf, Smola, & Müller, 1998), and semi-supervised learning (Chapelle, Schölkopf, & Zien, 2006; Xu, Jin, Zhu, King, & Lyu, 2008; Xu, Zhu, Lyu, & King, 2007; Zhu, Kandola, Ghahramani, & Lafferty, 2005). We focus here on kernel methods for supervised learning, where the basic idea is to use the so-called kernel trick to implicitly map the data from the ordinal input space to a high dimensional feature space, in order to make the data more separable. Usually, the aim of kernel-based classifiers is to find an optimal linear decision function in the feature space, based on certain criteria. The optimal linear decision hyperplane could be, for example, the one that can maximize the margin between two different classes of data (as used in the Support Vector Machine (SVM) (Vapnik, 1998)), or the one that minimizes the within-class covariance and at the same time maximizes the between-class covariance (as used in the Kernel Fisher Discriminant Analysis (KFDA) (Mika et al., 1999, 2003)), or the one that minimizes the worst-case accuracy bound (as used in the Minimax Probability Machine (Huang, Yang, King, & Lyu, 2004; Huang, Yang, King, Lyu, & Chan, 2004; Huang, Yang, Lyu, & King, 2008; Lanckriet, Ghaoui, Bhattacharyya, & Jordan, 2002)).

These kernel methods usually achieve higher prediction accuracy than their linear forms (Schölkopf & Smola, 2002). The reason is that the linear discriminant functions in the feature space can represent complex separating surfaces when mapped back to the original input space. However, one drawback of standard SVM is that it does not consider the data distribution and cannot properly output the probabilities or confidences for the resultant classification (Platt, 1999; Wu, Lin, & Weng, 2004). One needs special transformation in order to output probabilities. Therefore, it takes a lot of extra effort in order to be applied in systems that contain inherent uncertainty. In addition, the linear discriminant function can only separate two classes. For multi-category problems, we may resort to approaches such as one-against-one or one-against-others to vote on which class should be assigned (Hsu & Lin, 2002).

One approach to obtaining classification probabilities is to use a statistical pattern recognition technique, in which the probability

* Corresponding author. Tel.: +852 2609 8398; fax: +852 2603 5024.
E-mail addresses: zlxu@cse.cuhk.edu.hk (Z. Xu), k.huang@bristol.ac.uk (K. Huang), jkzhu@cse.cuhk.edu.hk (J. Zhu), king@cse.cuhk.edu.hk (I. King), lyu@cse.cuhk.edu.hk (M.R. Lyu).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.11.005

density function can be derived from the data. Future items of data can then be classified using a Maximum A Posteriori (MAP) method (Duda, Hart, & Stork, 2000). One typical probability estimation method is to assume multivariate normal density functions over the data. The multivariate normal density functions are easy to handle; moreover some problems can also be regarded as Gaussian problems if there are enough examples, although in practice the Gaussian distribution cannot be easily satisfied in the input space.

To solve these problems, in this paper we propose a Kernel-based Maximum A Posteriori (KMAP) classification method under a Gaussianity assumption in the feature space. With this assumption, we derive a non-linear discriminant function in the feature space, in contrast to current kernel-based discriminant methods that rely only on using an assumption of linear separability for the data. Moreover, the derived decision function can output the probabilities or confidences. In addition, the distribution can be very complex in the original input space when it is mapped back from the feature space. This is analogous to the case in which a hyperplane derived with KFDA or SVM in the feature space could lead to a complex surface in the input space. Therefore, this approach sets a more valid foundation than the traditional multivariate probability estimation methods that are usually conducted in the input space.

Generally speaking, distributions other than the Gaussian function can also be assumed in the feature space. However, under a distribution with a complex form, it is hard to get a closed-form solution and easy to over-fit. More importantly, with the Gaussian assumption, a kernelized version can be derived without knowing the explicit form of the mapping functions for our model, while it is still difficult to formulate the kernel version for other complex distributions.

It is important to relate our proposed model to other probabilistic kernel methods. Kernel-based exponential methods (Canua & Smola, 2006) use parametric exponential families to explicitly build mapping functions from the input space to the feature space. It is also interesting to discuss the Kernel Logistic Regression (KLR) (Zhu & Hastie, 2005), which employs the logistic regression to estimate the density function and still leads to a linear function in the kernel-induced feature space. The kernel-embedded Gaussian mixture model in Wang, Lee and Zhang (2003) is related to our model in that a similar distribution is assumed, but their model is restricted to clustering and cannot be directly used in classification.

The appealing features of KMAP are summarized as follows. First, one important feature of KMAP is that it can be regarded as a more generalized classification model than KFDA and other kernel-based algorithms. KMAP provides a rich class of family of kernel-based algorithms, based on different regularization implementations. Another important feature of KMAP is that it can output the probabilities of assigning labels to future data, which can be seen as the confidences of decisions. Therefore, the proposed method can also be seen as a Bayesian decision method, which can further be used in systems that make an inference under uncertainty (Smith, 1988). Moreover, multi-way classification is as easy as binary classification in this model because only probability calculation is involved and no one-against-one or one-against-others voting is needed. As shown in Section 2.4, KMAP has the time complexity $\mathcal{O}(n^3)$ (where n is the cardinality of data), which is in the same order as that of KFDA. In addition, the decision function enjoys the property of sparsity: only a small number of eigenvectors are needed for future prediction. This leads to low storage complexity.

The proposed algorithm can be applied in many pattern recognition tasks, e.g., face recognition, character recognition, and others. In order to evaluate the performance of our proposed method, extensive experiments are performed on eight benchmark data sets from the UCI repository and on three standard face data sets. Experimental results show that our proposed method achieves very competitive performance on UCI data. Moreover, its advantage is especially prominent in face data sets, where only a small amount of training data are available.

The remainder of this paper is organized as follows. In Section 2, we derive the kernel-based MAP classification model in the feature space and discuss the parameter estimation techniques. Then the kernel calculation procedure and the theoretical connections between the KMAP model and other kernel methods are discussed. Section 3 first reports the experiments on UCI data sets against other competitive kernel methods, then evaluates our model's performance on face data sets. Section 4 draws conclusions and lists possible future research directions.

We use the following notation. Let $\mathcal{X} \in \mathbb{R}^d$ denote the original d-dimensional input space, where an instance $\mathbf{x}$ is generated from an unknown distribution. Let $\mathbb{C} = \{1, 2, \ldots, m\}$ be the set of labels where m is the number of classes. Let $P(C_i)$ denote the prior probability of class C_i. Let n_i be the number of observed data points in class C_i and n be the amount of training data. A Mercer kernel is defined as a symmetric function κ, such that $\kappa(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, where Φ is a mapping from $\mathcal{X}$ to a feature space $\mathcal{H}$. The form of kernel function κ could be a linear kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, a Gaussian RBF kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/\sigma^2)$, or a polynomial kernel function, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^p$, for some σ and p respectively. A kernel matrix or Gram matrix $G \in \mathbb{R}^{n \times n}$ is a positive semi-definite matrix such that $G_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$. G can be further written as $[G^{(1)}, G^{(2)}, \ldots, G^{(m)}]$, where $G^{(i)}$ is an $n \times n_i$ matrix and denotes the subset of G relevant to class C_i. The covariance matrix of $G^{(i)}$ is denoted by $\Sigma_{G^{(i)}}$. We denote μ_i and Σ_i as the mean vector and covariance matrix of class C_i in the feature space, respectively. The set of eigenvalues and the set of eigenvectors belonging to Σ_i are represented as Λ_i and Ω_i. We write $p(\Phi(\mathbf{x})|C_i)$ as the probability density function of class C_i.

2. Kernel-based maximum a posteriori classification

In contrast with the assumption of traditional MAP algorithms, that the data points satisfy multivariate normal distribution in the input space, we assume that the mapped data in the high dimensional feature space follow such a distribution. This is meaningful in that the distribution can be very complex in the original input space when the Gaussian distribution is mapped back from the kernel-induced feature space. In the same sense, the decision boundary can be more complex when the quadratic decision boundary is projected into the input space.

In order to make a clear illustration of the reasonability of the Gaussian distribution in the kernel-induced feature space, two synthetic data sets, **Relevance** and **Spiral**, are used in this paper. We draw the decision boundary of discriminant functions conducted in the input space and the feature space, respectively. **Relevance** is a data set where only one dimension of the data is relevant to separate the data. **Spiral** can only be separated by highly non-linear decision boundaries. Fig. 1 plots the boundaries of the discriminant functions for the traditional MAP algorithm and the kernel-based MAP algorithm on these two data sets.

It can be observed that the MAP classifier with the Gaussian distribution assumption in the kernel-induced feature space always produces more reasonable decision boundaries. For **Relevance** data, a simple quadratic decision boundary in the input space cannot produce good prediction accuracy. However, the kernel-based MAP classifier separates these two classes of data smoothly. The difference between the boundaries of these two algorithms is especially significant for **Spiral**. This indicates that the kernel-based MAP classification algorithm can better fit the distribution of data points through the kernel trick.