Chaos, Solitons and Fractals 89 (2016) 175–186

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

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

Multiple kernel boosting framework based on information measure for classification

Chengming Qi[a,b,*], Yuping Wang[a], Wenjie Tian[a], Qun Wang[b]

[a] College of Automation, Beijing Union University, Beijing 100101, China
[b] School of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China

ARTICLE INFO

Article history:
Available online 18 December 2015

Keywords:
Classification
Ensemble learning
Kullback–Leibler distance
Multiple kernel learning (MKL)

ABSTRACT

The performance of kernel-based method, such as support vector machine (SVM), is greatly affected by the choice of kernel function. Multiple kernel learning (MKL) is a promising family of machine learning algorithms and has attracted many attentions in recent years. MKL combines multiple sub-kernels to seek better results compared to single kernel learning. In order to improve the efficiency of SVM and MKL, in this paper, the Kullback–Leibler kernel function is derived to develop SVM. The proposed method employs an improved ensemble learning framework, named KLMKB, which applies Adaboost to learning multiple kernel-based classifier. In the experiment for hyperspectral remote sensing image classification, we employ feature selected through Optional Index Factor (OIF) to classify the satellite image. We extensively examine the performance of our approach in comparison to some relevant and state-of-the-art algorithms on a number of benchmark classification data sets and hyperspectral remote sensing image data set. Experimental results show that our method has a stable behavior and a noticeable accuracy for different data set.

1. Introduction

Kernel-based methods [1], such as support vector machine (SVM), kernel principal component analysis (KPCA) and kernel Fisher discriminant analysis (KFDA), have been widely used to solve some machine learning problems in the past several decades. SVM, one of the most successful applications in kernel-based methods, has shown to be powerful tools for solving various problems in machine learning and data mining community. These methods are based on mapping data from the input feature space to a kernel feature space of higher dimensionality, where even linear methods can deliver very impressive performance. The mapping is determined implicitly by a kernel function, which computes the inner product of data points in a feature space.

Despite the success of kernel-based methods, a poor kernel can lead to impaired prediction performance. Choosing and/or constructing the appropriate kernel function and appropriate feature space is crucial for achieving good performance. To select the optimum kernel function for SVM, in this paper, we construct a kernel function which is constructed using the Kullback–Leibler divergence, substituting to the Euclidean distance in the Gaussian kernel. Kullback–Leibler divergence [2] (also relative entropy, or KL divergence) is an information measure of the difference between two probability distributions in probability theory and information theory. The detail is described in Section 4.

Multiple kernel learning (MKL) [3] is a powerful field of machine learning. MKL aims at learning an optimal combination of a set of predefined base kernels and finds an automatic combination of kernel functions. Compared with traditional single fixed kernel methods, MKL does

* Corresponding author at: College of Automation, Beijing Union University, Beijing 100101, China. Tel.: +86 15327217430.
E-mail address: qichengming111@sina.com (C. Qi).

http://dx.doi.org/10.1016/j.chaos.2015.10.024
0960-0779/© 2015 Published by Elsevier Ltd.

exhibit its flexibility of automated kernel learning, and also reflect the fact that typical learning problems often involve multiple, heterogeneous data sources. A great deal of analysis and algorithms for MKL focus on learning finite linear combinations of given base kernels. The idea of MKL can be applied to all sorts of kernel-based classifiers, such as SVM and KFDA, leading to SVM-based MKL and discriminant MKL, respectively. Our work in this paper only focus on SVM-based MKL formulations.

To evaluate the performance of our method, we apply our method to publicly available data sets and hyperspectral image (HSI) classification. A HSI can be viewed as an image cube where the first two dimensions indicates the spatial coordinate of the image and the third represents the number of bands of the image. Due to the availability of a large number of bands, "the curse of dimensionality" [4] and computation complexity are become two critical issues for HSI classification. This results in high redundancy between the spectral bands can lead to poor generalization capabilities of the classifier. However, it can be avoided by using Dimension Reduction (DR). A highly used strategy of DR is feature extraction or feature selection techniques [5]. To reduce the dimensionality of HSI, in this paper, we employ the Optional Index Factor (OIF) [6] to select the most informative and the least correlative bands for classification. The OIF is an unsupervised method that takes the bands' correlation into consideration and resort to searching the bands combination with maximum information. High value of OIF indicates the optimum combination of bands out of all possible 3-band combinations.

In order to improve the efficiency of SVM and MKL, in this paper, we construct a Kullback–Leibler kernel function to develop SVM, and employ an improved ensemble learning to propose an multiple kernel boosting framework for classification, named KLMKB. In comparison to other ensemble classifiers, three specific contributions of the our method can be summarized as follows: (1) we present a novel framework based on information measure for multiple kernel boosting, which applies the Kullback–Leibler distance to construct SVM kernel function; (2) we conduct experiments on publicly available benchmark dataset and hyperspectral image for validating the performance of our method by comparing with various state-of-the-art algorithms; and (3) we evaluate various parameters of our method and attempt to provide a trade off between accuracy and efficiency.

The remainder of this paper is organized as follows. In Section 2, we review the related work. Next in Section 3, we introduce the preliminaries, including SVM, MKL and OIF. Section 4 formulates the proposed framework of KLMKB. We show the experimental results and evaluate various parameters of our method in Section 5, and conclude in Section 6.

2. Related work

As a promising data mining approach, SVM is a margin-based discriminative classifier and based on the principles of structural risk minimization [7] which minimizes the probability of misclassifying a data point drawn randomly from an unknown probability distribution. The standard SVM only utilizes a single kernel function with fixed parameters, which necessitates model selection for good classification performance. Kernel-based methods have been widely used to solve some machine learning problems in the past several decades [8–10].

Single kernel learning usually needs to choose proper kernel parameters, while MKL usually searches for linear/nonlinear combination of predefined base kernels by maximizing the margin maximization. MKL provided more flexibility in solving similarities of data source than single kernel learning. The MKL framework learned both the optimal weights for combining the kernels and the SVM solution in a joint optimization problem, in which the objective functions were built in terms of the SVM structural risk functional.

Usually, MKL was formulated as a semidefinite programming (SDP) [11] or a second-order cone programming problem (SOCP) [12]. However, due to the high computational cost of solving those programming problems, this class of MKL only handle small-scale or medium-scale datasets. To address large scale kernel learning, various methods were developed. Rakotommonjy et al. [13] proposed SimpleMKL where the kernel weights are obtained by a reduced gradient descent method. Furthermore, semi-infinite linear programming (SILP) [14], sparse MKL [15] and SpicyMKL [16] were proposed to solve MKL problems. However, solving such joint optimization problem is far more complex than training a SVM classifier. Recently, Cortes [17–19] proposed two-stage procedure to address the problem. The first stage finds the optimal weights to combine the kernels, which makes use of the information from the complete training data and can be computed efficiently, and the second stage trains a standard SVM using the combined kernel.

There are many methods to achieve classifier diversity. The most popular method is to use different training datasets to train individual classifiers. Such datasets are often obtained through re-sampling techniques. To improve the limited classification performance of the real SVM, SVM ensemble with bagging (bootstrap aggregating) or boosting is proposed. In both bagging and boosting, the trained individual SVMs are aggregated to make a collective decision. SVM ensemble is essentially a type of cross-validation optimization of single SVM, having a more stable classification performance than other models. In bagging, each individual SVM is trained independently using the randomly chosen training samples via a bootstrap technique. In boosting, each individual SVM is trained using the training samples chosen according to the sample's probability distribution, such as, Boost-SMO [20], Boost-SVM [21], AdaBoost with SVM [22,23]. Various simulation results for hyperspectral remote sensing data show that the SVM ensemble with bagging or boosting greatly outperforms a single SVM in terms of classification accuracy [11,24].

Hyperspectral sensors divide the electromagnetic spectrum into hundreds of spectral bands, which can provide the potential and detailed land-cover distinction and identification [25]. Among Hyperspectral images applications, classification is one of the most important tasks for successful data exploitation. Classification of HSI consists of sequential steps such as pre-processing, feature extraction, feature selection, segmentation, classification, and post-processing. Those steps have been accomplished significant progress. However, the extremely high dimensionality of these data