Optik 125 (2014) 1024–1029

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A framework for classification with single feature kernel matrix

Hailong Huang*, Xin Zuo, Chao Huang, Jianwei Liu

Research Institute of Automation, China University of Petroleum, Beijing 102249, China

ARTICLE INFO

Article history:
Received 20 March 2013
Accepted 21 July 2013

Keywords:
Kernel matrix
Support Vector Machine (SVM)
Prediction error rate
Support vector

ABSTRACT

In this paper, we propose a novel classification framework using single feature kernel matrix. Different from the traditional kernel matrices which make use of the whole features of samples to build the kernel matrix, this research uses features of the same dimension of any two samples to build a sub-kernel matrix and sums up all the sub-kernel matrices to get the single feature kernel matrix. We also use single feature kernel matrix to build a new SVM classifier, and adapt SMO (Sequential Minimal Optimization) algorithm to solve the problem of SVM classifier. The results of the experiments on several artificial datasets and some challenging public cancer datasets display the classification performance of the algorithm. The comparisons between our algorithm and L_2-norm SVM on the cancer datasets demonstrate that the accuracy of our algorithm is higher, and the number of support vectors selected is fewer, indicating that our proposed framework is a more practical approach.

1. Introduction

The study of machine learning methods based on the input and the output data is one of the most important topics in modern intelligence technology. These methods are designed to find the rules from the observation data and employ the rules to predict the label of the future sample. There are two kinds of datasets: one is large samples and the other is small samples. Statistic theories have proved to be applicable for the cases with large samples, yet do not perform well for some particular cases with small samples. To compensate the deficiency, Vapnik proposed statistic learning theory [1]. Based on this new theory, SVM has been developed, and it is considered to be a very popular and successful example in machine learning.

The good performance of SVM in many applications benefits from the usage of many key technologies such as the largest margin hyperplane, convex quadratic programming, nonlinear kernel mapping, slack variables, and sparse solutions. Constructing a kernel function which can help improve prediction performance is a major concern in the academic world.

For datasets with complex structure, the accuracy of linear classifier is poor, thus nonlinear mapping function to implicitly map the input data points from the input space to a possibly high dimensional nonlinear space has been introduced. In the high dimensional space, linear classifier can achieve good performance. Kernel methods have some excellent characteristics:

(1) Kernel functions can avoid the problem of dimension disaster. The dimension of the kernel matrix is the same as the number of samples, thus can decrease the computation complexity. Therefore, kernel functions can cope with datasets with high dimension effectively.
(2) There is no need to know the explicit expression of nonlinear function.
(3) The type and parameters of kernel functions can implicitly affect the relationship between the input space and high dimensional space, then determine the properties in high dimensional space, and change the ability of kernel machines.

In the field of kernel machine researches, the most focused area is the determination of the best kernel function, and the values of its parameters. And many researches used all the features of samples to build the kernel matrix. In this article, we propose a new kind of method to build the kernel matrix, which uses only one single feature to build the sub kernel matrix, and then uses the sub kernel matrices to build the kernel matrix. Our new kernel matrix inherits the above characteristics but differs from the traditional methods. In the traditional methods, the components of the kernel matrix are calculated by two samples through kernel

* Corresponding author at: 260 mailbox, China University of Petroleum, Changping District, Beijing 102249, China. Tel.: +86 010 89733306; fax: +86 010 89731185.
E-mail address: huanghailong3520@163.com (H. Huang).

0030-4026/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijleo.2013.07.104

function. It distinguishes samples but does not consider the features. In our new method, we use only one feature to build the sub kernel matrix every time, and use all the sub kernel matrices to build the kernel matrix. We not only distinguish different samples but also features. Experimental results on artificial and real world datasets show that this new method provides a good prediction of generalization.

2. L$_2$-norm SVM

Suppose that the training dataset consists of a set of m samples $\{x_i, y_i\}_{i=1}^{m}$ where $x_i \in R^n, y_i \in \{-1, 1\}$. The original problem of SVM can be expressed as the following

$$\begin{aligned}&\min_{w,b,\xi} \frac{1}{2} w^T w + c \sum_{i=1}^{m} \xi_i \\ &s.t. y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, 2, \ldots, m\end{aligned} \tag{1}$$

Based on the standard L$_2$-norm SVM, researchers proposed some other formats, such as the Least Square SVM. From problem (1), we can get a linear classifier. However, for some datasets with complex inner structures, researchers introduced nonlinear function into problem (1) and we got

$$\begin{aligned}&\min_{w,b,\xi} \frac{1}{2} w^T w + c \sum_{i=1}^{m} \xi_i \\ &s.t. y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, 2, \cdots, m\end{aligned} \tag{2}$$

where $\phi()$ is the nonlinear function, and c is a non-negative regularization parameter fixed by the user which determines the tradeoff between the maximum margin and the minimum empirical risk.

Problem (2) is called the primal optimization problem. Introducing the Lagrange multipliers we get the dual optimization problem

$$\begin{aligned}&\min_{\alpha} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) - \sum_{i=1}^{m} \alpha_i \\ &s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, 0 \leq \alpha_i \leq c, \quad i = 1, \ldots, m\end{aligned} \tag{3}$$

The traditional kernel matrix is defined as

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{4}$$

which makes use of the inner production of two samples in high dimensional space.

Problem (3) is a quadratic programming problem, and we can use the kernel matrix to build the quadratic gram matrix

$$Q_{ij} = y_i y_j K(x_i, x_j) \tag{5}$$

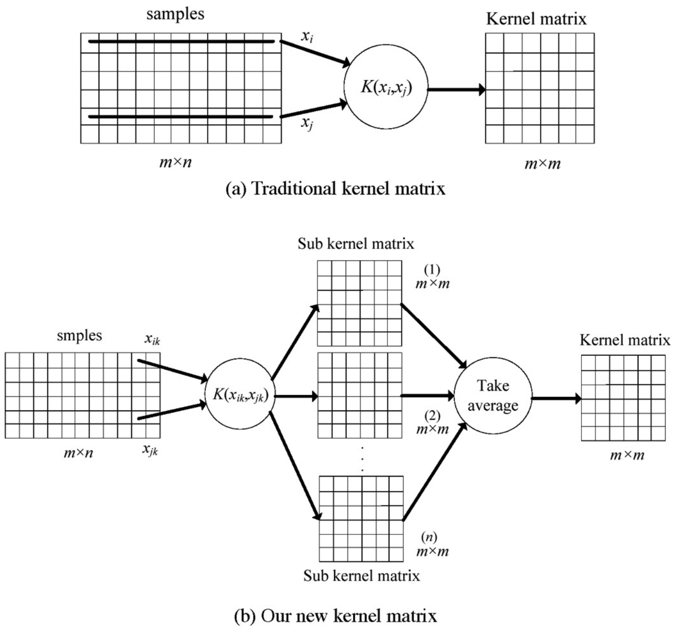
Then, problem (3) can be reformulated

$$\begin{aligned}&\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\ &s.t. y\alpha^T = 0, 0 \leq \alpha_i \leq c, \quad i = 1, 2, \ldots, m\end{aligned} \tag{6}$$

where $e = [1, 1, \cdots, 1]^T$.

The discrimination function based on this kind of kernel matrix is

$$sgn(w^T \phi(x) + b) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b\right). \tag{7}$$



(a) Traditional kernel matrix

(b) Our new kernel matrix

Fig. 1. Schematic diagrams of two different methods to build the kernel matrix. (a) Traditional kernel matrix and (b) sub kernel matrix.

3. SVM based on single feature kernel matrix

3.1. Traditional kernel matrix

Our new kernel matrix is inspired by RBF kernel matrix. The components of RBF kernel matrix can be calculated as follows.

$$\begin{aligned}K(x_i, x_j) &= \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{(x_{i1} - x_{j1})^2}{2\sigma^2}\right) \exp\left(-\frac{(x_{i2} - x_{j2})^2}{2\sigma^2}\right) \cdots \exp\left(-\frac{(x_{in} - x_{jn})^2}{2\sigma^2}\right) \\ &= K(x_{i1}, x_{j1}) K(x_{i2}, x_{j2}) \cdots K(x_{in}, x_{jn}) \\ &= \prod_{k=1}^{n} K(x_{ik}, x_{jk})\end{aligned}$$

The RBF kernel matrix is expressed as the product $K(x_{ik}, x_{jk})$ Our new method takes the element of kernel matrix as the average of $K(x_{ik}, x_{jk})$. We will derive the formulations below.

3.2. Single feature kernel matrix method

Firstly, we define single feature kernel matrix including sub kernel matrix and kernel matrix.

Sub kernel matrix can be got through (8), which used only one feature each time, and the average of all the sub kernel matrices gives rise to the kernel matrix as (9):

$$K_k^{new}(x_{ik}, x_{jk}) \overset{def}{=} \phi(x_{ik}) \phi(x_{jk}), \quad k = 1, 2, \ldots, n \tag{8}$$

$$K^{new} \overset{def}{=} \frac{1}{n} \sum_{k=1}^{n} K_k^{new} \tag{9}$$

We compare our new kernel matrix with traditional kernel matrix in Fig. 1.