Information Sciences 322 (2015) 133–149

Contents lists available at ScienceDirect

Information Sciences

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

An efficient Gaussian kernel optimization based on centered kernel polarization criterion

Meng Tian [a,b], Wenjian Wang [a,*]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, PR China
[b] School of Science, Shandong University of Technology, Zibo 255049, PR China

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form 15 April 2015
Accepted 11 June 2015
Available online 17 June 2015

Keywords:
Gaussian kernel
Kernel selection
Pattern classification
Centered kernel polarization
Local multiclass centered kernel polarization

ABSTRACT

The success of kernel-based learning methods is heavily dependent on the choice of a kernel function and proper setting of its parameters. In this paper, we optimize the Gaussian kernel for binary-class problems by using centered kernel polarization criterion. This criterion is an extension of kernel polarization and a simplified style of centered kernel alignment. Compared with formulated kernel polarization criterion, the proposed criterion has a defined geometrical significance, and it can locate the global optimal point with less influence of threshold selection. Furthermore, the approximate criterion function can be proved to have a determined global minimum point by adopting the Euler–Maclaurin formula under weaker conditions. In addition, taking the preservation of within-class local structure into account, we present an evaluation criterion named local multiclass centered kernel polarization in multiclass classification scenario. Comparative experiments are conducted on some benchmark examples with three Gaussian kernel based learning methods and the results well demonstrate the effectiveness and efficiency of the proposed quality measures.

1. Introduction

Kernel-based learning methods, such as support vector machine (SVM) [27], kernel principal component analysis (KPCA) [23]. and kernel linear discriminant analysis (KLDA) [19], provide high performance for solving a wide range of different problems in machine learning community. These methods work by mapping the input data into a high-dimensional feature space and then build linear algorithms in the feature space to implement nonlinear counterparts in the input space. The key to the success of kernel methods is the incorporation of the "kernel trick" which computes a kernel function as the inner product between each pair of points in the feature space without computing their images directly. Thus these kernel methods combine the advantages of linear and non-linear classifiers in terms of efficient training time, elegant compatibility with high-dimensional data.

It is reasonable to hope that the mapped classes in the feature space possess a better linear separability compared with that obtained in the input space for a classification task. However, the classification performance of kernel methods can be even worse than that of their linear counterparts in the original input space when the kernel functions are not well chosen [33]. So whether kernel methods behave well largely depends on their adopted kernel functions. It is well known that the choice of the kernel function is a challenging problem.

* Corresponding author. Tel.: +86 351 7017566; fax: +86 351 7018176.
E-mail addresses: luckywalter@163.com (M. Tian), wjwang@sxu.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.ins.2015.06.010
0020-0255/

In the literatures, kernel selection is usually tackled by cross validation and leave-one-out method. These two methods are data-independent, but they both suffer heavy computational complexity. To remedy this problem, in the context of SVM, some upper bounds on the generalization error have been proposed [6,7,11]. Of these bounds, the radius-margin bound is most commonly used in practice. However, it still requires the whole learning process for evaluation like cross validation and leave-one-out method.

In order to obtain a better computation efficiency, many universal data-dependent kernel evaluation measures have been derived by optimizing the measure of data separation in the feature space. Based on Fisher discrimination criteria, Refs. [28,31,33] proposed different approaches to optimize the kernel parameters. However, the use of Fisher criteria tends to give undesired results if samples in some class form several separative clusters, especially for the case of multimodally distributed data [25]. By using the measure called "alignment", Ref. [9], for the first time proposed a kernel target alignment criterion to optimize the kernel function. This criterion can measure the similarity between two kernel matrices or the degree of agreement between a kernel and a given target function. Beginning with kernel target alignment, many measurement criteria have been derived for kernel selection, such as kernel polarization [2], feature space based kernel matrix evaluation measure [20] and local kernel polarization [29].

Basically, kernel target alignment is the most commonly used efficient kernel measure criterion. Some researchers found that the sensitivity of kernel target alignment in case of uneven class distribution will drop drastically [14]. Refs. [8,20] showed a kernel matrix with a low kernel alignment value may have a very good performance. This means having a very high kernel target alignment is only a sufficient condition, but not a necessary condition, for kernel function to be a good one for a given task [20]. Therefore, Ref. [8] proposed a new criterion, centered kernel alignment, to modify kernel target alignment by adopting the notion of centering in the feature space. In addition to giving a simple concentration bound for centered kernel alignment, the existence of good predictors for kernel with high alignment both for classification and for regression has been shown. By this criterion, a steepest ascent approach based on forward stagewise additive method has been presented for multiple kernel learning. The approach achieves good performance across a variety of real-world data sets without discretizing the space of base kernels [1]. Multiple kernel clustering based on centered kernel alignment has also been proposed [18].

Recently Ref. [34] proposed an efficient Gaussian kernel optimization method, which works by maximizing the formulated kernel target alignment (in fact, it is the formulated kernel polarization). The contribution of this work lies in obtaining a differentiable objective function having a determined minimum point. More remarkably, the approximate analytical solution of the formulated criterion can be obtained by using the Euler–Maclaurin formula. Furthermore, the optimization has been solved with high computation efficiency by using a Newton-based algorithm with a unique starting point to locate the best local minimum compared with the searching procedure in [28]. However, the objective function curve of alignment value depending on the kernel parameter on some data sets monotonically increases very slowly when the parameter is greater than the optimal parameter, and then the selected parameter may be dependent on the threshold values of the search algorithm. Besides, the proof of having a determined global minimum point for approximate formulated criterion was obtained under strong constraint conditions.

We propose an effective surrogate measure based on kernel polarization, namely, centered kernel polarization. The approximate criterion function can be proved to have a determined global minimum point for two-class pattern classification tasks under weaker constraint conditions than those in [34]. We note that the proposed criterion is similar to the Hilbert–Schmidt Independence Criterion (HSIC) [13], which is a practical criterion for independence test in the context of independent component analysis (ICA). In this paper, we mainly tune the Gaussian kernel parameter on the basis of centered kernel polarization, and study the analytic properties and geometrical significance of the proposed criterion as well. In addition, based on the works in [29,30], we put forward a new multiclass evaluation criterion named local multiclass centered kernel polarization by taking the local structure preservation into account.

The rest of this paper is organized as follows. Section 2 gives a short description of some properties of Gaussian kernel and three criteria, namely, kernel target alignment, kernel polarization and centered kernel alignment. Section 3 discusses the continuous differentiability of the formulated centered kernel polarization and proves that the approximate criterion function has a determined global minimum point. In addition, by exploring the relationship among the centered kernel polarization criterion and two other off the shelf kernel evaluation measures, the geometric meaning of the proposed criterion is revealed. Section 4 describes the proposed local multiclass evaluation criterion in detail. Experimental results are presented in Section 5.

In this paper, all analyses are based on Gaussian kernel function. In the following, K denotes a kernel function, capital-case boldface symbols are used for matrices, $< \cdot, \cdot >$ denotes a dot product, and $< \cdot, \cdot >_F$ denotes a Frobenius inner product.

2. Preliminaries

2.1. The Gaussian kernel for classification

Recently, the use of kernel functions in machine learning and data mining community has received considerable attention. The kinds of kernel K we will be interested in are such that for all samples x_i and x_j, where $x_i, x_j \in \mathcal{X} \subset \mathbb{R}^m$, and $\mathcal{X}$ is the input space: