Pattern Recognition 68 (2017) 38–51

Contents lists available at ScienceDirect

Pattern Recognition

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

Success based locally weighted Multiple Kernel combination

Raghvendra Kannao*, Prithwijit Guha

Indian Institute of Technology Guwahati, Guwahati, Assam, 781039, India

ARTICLE INFO

Article history:
Received 10 June 2016
Revised 14 January 2017
Accepted 23 February 2017
Available online 28 February 2017

Keywords:
Support vector machine
Multiple kernel learning
Kernel alignment
Localized multiple kernel learning
Regions of success
Success prediction functions
Feature selection
Kernel selection
Feature fusion
Support vector regression

ABSTRACT

Multiple Kernel Learning (MKL) literature has mostly focused on learning weights for base kernel combiners. Recent works using instance dependent weights have resulted in better performance compared to fixed weight MKL approaches. This may be attributed to the fact that, different base kernels have varying discriminative capabilities in distinct local regions of input space. We refer to the zones of classification expertize of base kernels as their "Regions of Success" (RoS). We propose to identify and model them (during training) through a set of instance dependent success prediction functions (SPF) having high values in RoS (and low, otherwise). During operation, the use of these SPFs as instance dependent weighing functions promotes locally discriminative base kernels while suppressing others. We have experimented with 21 benchmark datasets from various domains having large variations in terms of dataset size, interclass imbalances and number of features. Our proposal has achieved higher classification rates and balanced performance (for both positive and negative classes) compared to other instance dependent and fixed weight approaches.

1. Introduction

Support Vector Machines (SVM, henceforth) have proven to be a successful tool for solution of a wide range of classification problems since their introduction in [1]. SVM learns a maximum margin discriminative hyperplane, which in turn implies good generalization capabilities on unseen data [2]. The core idea of SVM is to formulate the margin maximization problem as a convex optimization problem, which has a single global minimum. SVM uses a kernel function for mapping the data to the "kernel space". A discriminative hyperplane is then learned in kernel space with the maximum margin criterion. Thus, selection of kernel function is a critical step in training SVM. Recent works have demonstrated the usability of weighted combination of multiple base kernels instead of a single one. In this work we introduce a new framework for instance dependent weighing of the kernels in the combination. We propose to link the weight assigned to a kernel to its performance in local regions of the feature sub-space. The motivation of our proposal is introduced in Section 1.3 before which, we briefly discuss the basic formulation of the SVM in Section 1.1 and the feature-kernel selection problem in Section 1.2.

1.1. The support vector machine

Given a set of n labeled training instances, $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ with $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, +1\}$, the hyperplane learned by SVM is given by, $f(\boldsymbol{x}_l) = \langle \boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x}_l) \rangle + b = 0$ where, the function $\boldsymbol{\Phi}(\cdot)$ maps the data from feature space ($\mathbb{R}^D$) to the kernel space ($\mathbb{R}^{D_\Phi}$). The hyperplane coefficient vector $\boldsymbol{w}$ and bias b are estimated by solving the following quadratic optimization problem.

$$\begin{aligned} &\textbf{Minimize} \quad \frac{1}{2} \parallel \boldsymbol{w} \parallel_2^2 + C \sum_{i=1}^{n} \zeta_i \\ &w.r.t. \quad \boldsymbol{w} \in \mathbb{R}^{D_\Phi}, \boldsymbol{\zeta} \in \mathbb{R}^n, b \in \mathbb{R} \\ &s.t. \quad y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i) \rangle + b) \geq 1 - \zeta_i \quad i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

Here, $\boldsymbol{\zeta}$ is a vector of slack variables and C controls the trade off between generalization (model simplicity) and classification error [2]. This optimization problem has one constraint per training instance and can be reformulated using its Lagrangian dual function as **Maximize** $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{l=1}^{n} \alpha_i \alpha_l y_i y_l \langle \boldsymbol{\Phi}(\boldsymbol{x_i}), \boldsymbol{\Phi}(\boldsymbol{x}_l) \rangle$ with respect to $\boldsymbol{\alpha} \in [0, C]^n$; such that, $\sum_{i=1}^{n} \alpha_i y_i = 0$ where, α_i is a Lagrange multiplier corresponding to the i^{th} constraint $y_i(\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i) \rangle + b) \geq 1 - \zeta_i$ in the primal formulation. This dual formulation is a typical quadratic programming problem (QP). Solving this gives us the hyperplane coefficient vector as $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\Phi}(\boldsymbol{x_i})$. Thus, the expression for the hyperplane can be

* Corresponding author.
E-mail addresses: raghvendra@iitg.ernet.in (R. Kannao), pguha@iitg.ernet.in (P. Guha).

http://dx.doi.org/10.1016/j.patcog.2017.02.029
0031-3203/

rewritten as,

$$f(\boldsymbol{x}_l) = \sum_{i=1}^{n} \alpha_i y_i \underbrace{\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_l) \rangle}_{\boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}_l)} + b = 0 \quad (2)$$

We note that only active constraints have non-zero values and the corresponding training instances are known as support vectors [2]. The label predicted by SVM for an unknown test pattern $\boldsymbol{x_l}$ is given by $sign(f(\boldsymbol{x_l}))$. SVMs use the inner product $\langle \boldsymbol{\Phi}(\boldsymbol{x_i}), \boldsymbol{\Phi}(\boldsymbol{x_l}) \rangle$ during training as well as testing. Hence, inner products of explicitly defined mapping functions $\boldsymbol{\Phi}(\cdot)$ can be replaced by a kernel function, $\boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}_l)$. The SVM hyperplane (Eq. 2) behaves as a linear classifier in the kernel space. Thus, the kernel function is entirely responsible for making data linearly separable in the kernel space [3]. Kernels compute the similarity (or dissimilarity) between two instances. The mappings defined by kernels are characterized by nature and distribution of input instances along with similarity functions used. However, various kernels with standard similarity functions have been used successfully in the literature. Some popularly used kernels are Linear ($\boldsymbol{k}_L(\boldsymbol{x}_i, \boldsymbol{x}_l) = \langle \boldsymbol{x}_i, \boldsymbol{x}_l \rangle$), Radial Basis Function (RBF) ($\boldsymbol{k}_R(\boldsymbol{x}_i, \boldsymbol{x}_l) = \exp(-\gamma(||\boldsymbol{x}_i - \boldsymbol{x}_l||_2^2)); \quad \gamma \in \mathbb{R}^+$) and Polynomial ($\boldsymbol{k}_P(\boldsymbol{x}_i, \boldsymbol{x}_l) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_l \rangle + 1)^q; \quad q \in \mathbb{N}$) [2].

Several works have used domain specific kernels (designed by experts) considering the nature of data. For example, string kernels are used for natural language processing and image classification [4]. Graph kernels [5] were used in drug discovery, chemo-informatics and bio-informatics for analyzing structural data. Aseervatham et al. [6] have used semantic kernel for classifying medical documents. The kernels suitable for a particular problem are generally specified by domain experts. However, the research community has mainly focused on automatic selection and learning of most suitable kernels for specific problems [3,7,8].

1.2. The feature and kernel selection problem

The SVM formulation as a quadratic optimization problem (Eq. 1) has a single global minimum and has implicit dependence on data dimension [2]. SVMs have overcome several problems of conventional discriminative models (neural networks, decision tree etc.) such as, convergence to local minima and explicit dependence on data dimension [1,2]. However, SVM faces the following three challenges. First, SVM testing becomes computationally expensive if the learned representation has a large number of support vectors. The number of support vectors depend on complexity of learned model and can be reduced substantially by optimal selection of features and kernels. Second, SVM has implicit dependence on data dimension. Thus, an external feature selection procedure is required to prevent degradation in performance of SVM due to presence of irrelevant features [3]. Third, various kernels have different notions of similarity and thus capture distinct views of features. Consequently, each kernel leads to a distinct hyperplane in feature space ($\mathbb{R}^D$). Thus, selection of optimal similarity function and its parameters is crucial for enhanced performance of SVM [2,3]. All these shortcomings of SVMs can be solved to an extent by an effective feature and (corresponding) similarity function selection procedure.

Several methods are proposed in literature for optimal selection of features and similarity functions (or kernels). A general purpose automatic feature selection procedure still remains an unsolved problem [3,7,8]. More recent works [3,7,9–11] have focused largely on learning data dependent kernels instead of selecting a particular one. These are learned by combining various standard as well as domain specific base kernels in multiple kernel learning (MKL) framework. These component kernels are known as base kernels. Each base kernel in the weighted combination may be defined either on the entire feature vector or on a subset of features. Thus, it is characterized by the subset of features used and the similarity function. The kernel combiners (weights) are learned from training data to determine the kernel's relevance to SVM decision. Most MKL methods have proposed to use fixed combiner weights [3,7,8] as opposed to instance dependent weighing approaches presented in [9,12]. It was observed that, the instance dependent kernel combination schemes worked better compared to the ones using fixed weights.

1.3. Proposed approach

Instance dependent MKL approaches learn a set of weighing functions instead of fixed weights. Given a test instance, weighing functions are used to determine the relevant base kernels. Existing instance dependent MKL approaches partition the feature space into non-overlapping regions. This partitioning is either guided by prior knowledge [12] or optimized to minimize classification error [9]. For each partitioned region, locally best performing kernel(s) is (are) identified and assigned high weight(s). The partitioned regions are then called as "regions of influence" [9] of best performing kernels. These approaches have inherently assumed the regions of influence of kernels to be linearly separable in feature space. On the other hand, we observe that each kernel has local regions of expertise in feature (sub)space, where it has good discriminative capability or high likelihood of successful classification (Fig. 1). Even so in most practical cases, these regions can not be obtained by merely partitioning the feature space with hyper-planes. This motivated us to link kernel weights to their discriminative capabilities in different local regions of feature space. We call these regions as **Regions of Success** (RoS) and thus, we name our proposal as **Success based Locally Weighted Kernel Combination** (**S-MKL**, henceforth).

We observe, SVMs using individual base kernels have diversity in classification errors if they are constructed using independent features and distinct similarity functions. This diversity in errors indicates that base kernels have expertise or discriminative capabilities in different regions of feature space (Fig. 1)[14,15]. This heterogeneity in discriminative regions may lead to superior performance subject to existence of suitable kernel combination schemes. However, fixed kernel weights may degrade the performance of kernel combination in local regions of the feature space if number of discriminative kernels are outnumbered by non-discriminative ones [9,12]. In S-MKL, we desire to suppress the non-discriminative kernels to effectively harness the error diversity in base kernels. This suppression ensures that the resultant kernel combination has higher discriminative capabilities leading to enhanced classification performance [3] and is more aligned with the "ideal kernel" [16,17].

We propose to identify and model the regions of success (RoS) of each base kernel in the feature space. SVMs are trained with individual base kernels and the correctly classified instances (both positive and negative) for each SVM are identified from the cross validation set. The (small spatial) neighborhoods of these correctly classified instances for a certain SVM collectively form the regions of success of the corresponding base kernel. We believe that base kernels have discriminative capabilities in their regions of success. We argue that in a linearly weighted combination of base kernels, it is necessary to promote the ones having discriminative capabilities while suppressing others. Such a weighing scheme aligns the combined kernel with the ideal one thereby leading to improved performance. This led us to propose an instance dependent weighing scheme in terms of **Success Prediction Functions** (SPFs, henceforth). Each base kernel has its own SPF and is learned through regression analysis over its RoS. For the SPF of a certain base kernel, target for regression model is set to 1.0 for correctly classified (by SVMs) training instances (from cross validation