Information Sciences 301 (2015) 124–140

Contents lists available at ScienceDirect

Information Sciences

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

Structural multiple empirical kernel learning

Zhe Wang *, Qi Fan, Sheng Ke, Daqi Gao *

Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 29 December 2014
Accepted 3 January 2015
Available online 9 January 2015

Keywords:
Multiple kernel learning
Empirical kernel mapping
Structural learning
Cluster information
Rademacher complexity
Pattern recognition

ABSTRACT

Multiple Kernel Learning (MKL) can boost classification performance through using multiple kernels rather than a single fixed one. Unlike the traditional MKL with the implicit kernels, Multiple Empirical Kernel Learning (MEKL) explicitly maps input data into multiple feature spaces. This paper focuses on MEKL and proposes an effective Threefold Structural MEKL (TSMEKL). The first fold structure is the space structural information between different mapped feature spaces. The second one is the class discriminant information within each mapped feature space. The third one is the cluster structural information of samples in each mapped feature space. The classical MEKL mainly pays attention to the first two structures, but neglects the last one. The proposed TSMEKL introduces the cluster structural information into MEKL. Doing so can simultaneously utilize the space, the class, and the cluster information in the way from globality to locality. Therefore, TSMEKL utilizes threefold structural information to result in the improvement of classification performance. To the best of our knowledge, it is the first time to introduce the cluster information into the MEKL framework. The main advantage of the developed TSMEKL is considering different folds of data information to improve classification performance. The experimental results validate the feasibility and effectiveness of TSMEKL. Moreover, we discuss the theoretical and experimental generalization risk bound of the proposed algorithm in terms of the Rademacher complexity.

1. Introduction

Unlike the traditional kernel-based methods which only use single fixed kernel [36,38,42,46,50], Multiple Kernel Learning (MKL) introduces multiple kernels into applications [3,9,53]. MKL is flexible and effective in depicting heterogeneous data sources. It is known that MKL maps the input sample x into different feature spaces F_ls, $\Phi_l(x): x \rightarrow \mathcal{F}_l,\ l = 1 \ldots M$ [3,30,43]. In practice, the mapping $\Phi(x)$ has two representations including the implicit and the explicit ones denoted by $\Phi^i(x)$ and $\Phi^e(x)$, respectively. The Implicit Kernel Mapping (IKM) $\Phi^i(x)$ is achieved through introducing a kernel function $k(x_i, x_j) = \Phi^i(x_i) \cdot \Phi^i(x_j)$, which implicitly determines the geometrical structure of the mapped data in feature spaces. Most of the existing MKL adopts IKM and thus is called Multiple Implicit Kernel Learning (MIKL). In contrast, the Empirical Kernel Mapping (EKM) $\Phi^e(x)$ can provide the explicit features of the x in the mapped Φ^e-space [40,55,56]. The MKL with the EKM is called as Multiple Empirical Kernel Learning (MEKL). Both IKM and EKM have their own characteristics. The $\Phi^i(x)$ is supposed to avoid the curse of dimensionality and simultaneously keep the linearity property of learning machines in the mapped space. The $\Phi^e(x)$ is supposed to be much easier in processing and analyzing the adaptability of kernels.

* Corresponding authors.
E-mail addresses: wangzhe@ecust.edu.cn (Z. Wang), gaodaqi@ecust.edu.cn (D. Gao).

http://dx.doi.org/10.1016/j.ins.2015.01.005
0020-0255/

Although MIKL has been widely applied [30,43,44], there is little attention on the variation of the mapping from the input space into the feature space [54]. Thus, in this paper we focus on MEKL. MEKL can be viewed as a data-dependent kernel learning technique since $\Phi^e(x)$ is directly generated based on the input data. The classical MEKL mainly introduces the existing algorithms into EKM and demonstrates the differences between EKM and IKM. However, the inherent characteristics of EKM are less discussed, which urges us to investigate the inherent structures of the feature spaces generated by multiple EKMs. Thus, here we give threefold structures including the space, the class, and the cluster in the feature spaces generated by multiple EKMs. The first fold structure is the space structural information between different mapped feature spaces. The second one is the class discriminant information within each mapped feature space. The third one is the cluster structural information of samples in each mapped feature space. Accordingly, we develop a novel Threefold Structural MEKL named TSMEKL. The proposed TSMEKL can simultaneously utilize the space, the class, and the cluster information in the way from globality to locality and improve classification performance. To the best of our knowledge, this work is the first time to introduce the cluster information to MEKL. The main contribution of the developed TSMEKL is to consider different fold structure information to boost classification performance.

In order to validate the effectiveness of the proposed TSMEKL, we employ our previous MEKL work called MultiK-MHKS [54] as the incorporated paradigm. MultiK-MHKS explicitly maps the input data into multiple feature spaces, each of which is expected to exhibit some geometrical structures of the original data from its own perspective. In doing so, all the feature spaces can complement the subsequent learning task. Moreover, MultiK-MHKS aims to make the projection of each feature space onto their corresponding basis vectors maximally close to the average projection of all feature spaces. Thus, MultiK-MHKS owns the space and the class structure information. Here, the proposed TSMEKL firstly captures the underlying structures distributing in the mapped samples by some unsupervised clustering techniques. Then, TSMEKL obtains multiple clusters to enclose the data of each class. Finally, TSMEKL introduces different clusters into the learning framework of MultiK-MHKS to achieve the threefold structure knowledge. The experimental results validate the feasibility and effectiveness of TSMEKL. We further discuss the theoretical and experimental generalization risk bound of TSMEKL in terms of the Rademacher complexity. It can be found that the advantage of TSMEKL not only inherits the structural information of both spaces and classes, but also adopts the cluster information, where the samples in the same cluster should be distributed as tightly as possible.

The rest of this paper is organized as follows. Section 2 reviews the related work of MKL. Section 3 gives the architecture of the proposed TSMEKL. Section 4 reports all the experimental results and gives further discussions on TSMEKL. Section 5 shows a theoretical and experimental generalization risk bound discussion on TSMEKL. Finally, the conclusions are presented in Section 6.

2. Related work

The classical MKL usually employs a linear convex combination of M kernels, i.e. $G = \sum_{l=1}^{M} \alpha_l K_l,\ \alpha_l \geqslant 0, \sum_{l=1}^{M} \alpha_l = 1$. In the literature [30], Lanckriet et al. introduce a convex Quadratically Constrained Quadratic Program (QCQP) to combine multiple kernels. To extend the method of Lanckriet et al. to large scale problems, Bach et al. [3] reformulate the QCQP as the Second-Order Cone Programming (SOCP) problem. Meanwhile, Sonnenburg et al. [43,44] reconstruct QCQP as the Semi-Infinite Linear Program (SILP) which recycles the standard Support Vector Machine (SVM) implementations [5,11,21]. Moreover, Jian et al. [24] address the issue of multiple kernels for Least Squares Support Vector Machine (LSSVM) by formulating a Semi-Definite Programming (SDP). Recently, researchers give some new formulations for multiple kernel optimization [15,45,34]. Hu et al. [20] introduce a sparse formulation into MKL to result in less support vectors. In their sparse formulation, the final ensemble kernel function can be calculated as a linear combination of kernels automatically. Yang et al. [58] discuss the problem of MKL in terms of searching for the optimal kernel weights. They employ the L_p-norm constraint onto the kernel weights in order to maintain all information of the base kernels. To group the features during the model development, Subrahmanya et al. [45] introduce a new Sparse MKL named SMKL for signal processing. SMKL is efficient in implementing the kernel-based methods by a convex primal formulation. Further, Schölkopf et al. [41] find that the Gram kernel matrix generated by all the training data does not always perform well when the matrix has large diagonal values. Hu et al. [20] apply a sparse formulation into MKL to generate a linear combination of multiple kernels, which significantly increases the running speed. Zhang et al. [60] discuss how to effectively map the input data into a high-dimension hidden space by a set of hidden non-linear functions. Their algorithm adopts different kinds of kernel functions and can result in a superior performance.

It can be found that the above MKL adopts IKM to implicitly represent the feature spaces. It is not necessary to show the exact forms of samples in the feature spaces since the geometrical structures are implicitly represented by the inner-product form $ker(x_i, x_j) = \Phi^i(x_i) \cdot \Phi^i(x_j)$. In contrast, EKM is easier to access and study the adaptability of the kernel for the input space than IKM [40,56]. Schölkopf et al. [40] discuss the relationship between the input and the feature spaces with the empirical and implicit kernels. They find that both IKM and EKM have the same geometrical structure. Kim et al. [25] denote that IKM has two problems. The one is the rising computational cost with respect to the increasing training size. The other is to recompute the whole decomposition when updating the eigenvectors with new data. However, EKM can avoid the above issues and be more flexible. Liang et al. [33] demonstrate that the separating hyperplane based on EKM is manipulable in the feature space. Further, Xiong et al. [56] derive an effective kernel optimization algorithm through employing a data-dependent kernel. Their idea is to maximize the class separability of the data in the empirical feature space. Wang et al.