Pattern Recognition 61 (2017) 197–209

Contents lists available at ScienceDirect

Pattern Recognition

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

MREKLM: A fast multiple empirical kernel learning machine

Qi Fan [a], Zhe Wang [a,c,*], Hongyuan Zha [b], Daqi Gao [a,*]

[a] Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai 200237, PR China
[b] School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, GA, United States
[c] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, PR China

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form
16 July 2016
Accepted 18 July 2016
Available online 20 July 2016

Keywords:
Multiple Kernel Learning
Empirical Kernel Mapping
Random projection
Analytical optimization
Classifier design
Pattern recognition

ABSTRACT

Multiple Empirical Kernel Learning (MEKL) explicitly maps samples into different empirical feature spaces in which the kernel features of the mapped samples can be directly provided. Thus, MEKL is much easier than the conventional Multiple Kernel Learning (MKL) in terms of processing and analyzing the structure of mapped feature spaces. However, the computational complexity of MEKL with M empirical feature spaces is $O(MN^3)$ where N is the number of training samples. The dimensions of the generated empirical feature spaces are approximate to N. When dealing with large-scale problems, MEKL cannot handle them properly due to the severe computation and memory burden. Moreover, most existing MEKL utilizes the gradient decent optimization to learn classifiers, but it is time consuming for training. Therefore, this paper proposes a Multiple Random Empirical Kernel Learning Machine (MREKLM) to overcome these problems. The proposed MREKLM adopts the random projection idea to map samples into multiple low-dimensional empirical feature spaces with lower computational complexity $O(MP^3)$, where $P(\ll N)$ is the number of the randomly selected samples. After that, MREKLM adopts an analytical optimization approach to directly deal with multi-class problems. The computational complexity of MREKLM is $O(M^3P^3)$. Experimental results also validate both efficiency and effectiveness of the proposed MREKLM. The contributions of this work are: (1) proposing a fast MEKL algorithm named MREKLM, (2) introducing an efficient random empirical kernel mapping approach, and (3) extending the capability of MEKL to handle large-scale problems.

1. Introduction

Kernel-based algorithms have been proven to have powerful performance for solving various problems in pattern recognition [3], image processing [10], data mining [34], and dimensionality reduction [36] community. Kernel-based algorithms map the samples from the input space $\mathcal{I}$ to the feature space $\mathcal{F}$ via a specific mapping function $\mathbf{\Phi}$, i.e. $\mathcal{I} \xrightarrow{\Phi} \mathcal{F}$, in which a relatively simple algorithm can result in impressive performance. As the mapping $\mathbf{\Phi}$ is determined implicitly by a kernel function, the selection of kernel is crucial for favorable performance. Unfortunately, to select an appropriate kernel is difficult for a specific problem. In recent years, instead of employing a single kernel, researchers propose Multiple Kernel Learning (MKL) [10] which adopts the combination of multiple candidate kernels, where the kernel weights are optimized during the training process. Compared with kernel-based algorithms employing a fixed kernel, MKL exhibits the flexibility in dealing with the problems involving multiple and heterogeneous data sources [35].

As multiple kernels are adopted, the main issue of MKL is how to combine them appropriately. Lanckriet et al. [27] introduce a convex combination of these kernels by constructing a convex Quadratically Constrained Quadratic Program (QCQP). Sonnenburg et al. [39] reformulate QCQP as a Semi-Infinite Linear Programming (SILP). Moreover, Jian et al. [18] address the issue of combining multiple kernels for Least Squares Support Vector Machine (LSSVM) by transforming MKL into a Semi-Definite Programming (SDP). Recently, Gönen et al. [16] propose a localized MKL composed of a kernel-based learning algorithm and a parametric gating model to assign local weight to each kernel, which are trained by a two-step alternating optimization algorithm. In the literature [2], Aiolli et al. demonstrate that the state-of-the-art MKL algorithms have the drawback that the time required to solve the associated optimization problem grows with the number of combined kernels. They propose a time and space efficient MKL named EasyMKL which can easily cope with hundreds of thousands of kernels and more. Alternatively, Wu et al. [45] propose a general strategy to pre-select a reasonable set of the candidate

* Corresponding authors at: Department of Computer Science & Engineering, East China University of Science & Technology, Shanghai 200237, PR China.
E-mail addresses: wangzhe@ecust.edu.cn (Z. Wang), gaodaqi@ecust.edu.cn (D. Gao).

http://dx.doi.org/10.1016/j.patcog.2016.07.027
0031-3203/

kernels before MKL optimization process based on the combination of minimal redundancy maximal relevance criteria and kernel target alignment.

Kernel mapping methods involved in the aforementioned algorithms are Implicit Kernel Mapping (IKM) which implicitly maps samples to feature spaces via the inner-product forms. It is the necessity of inner-product in IKM that restricts other methods unsatisfying the formulation to be kernelized. For instance, to formulate Kernel Direct Discriminant Analysis (KDDA) [33] is pretty difficult. For some linear discriminant algorithms, such as Orthogonal Linear Discriminant Analysis (OLDA) [47] and Uncorrelated Linear Discriminant Analysis (ULDA) [48], to directly kernelize them via the kernel trick is impossible, since they need to compute the singular value decomposition. Thus, in order to address the restriction of IKM, researchers introduce Empirical Kernel Mapping (EKM) [46] which maps samples into feature spaces via an explicitly represented mapping. With EKM, most algorithms can be kernelized directly due to the explicit representation of the mapped feature vectors, which results in much easier processing and analyzing the structure of the generated empirical feature spaces. To distinguish from the conventional MKL, the MKL adopting EKM to construct empirical feature spaces is denoted as Multiple Empirical Kernel Learning (MEKL).

However, the conventional MEKL suffers from the following four shortcomings. Firstly, the computational complexity for constructing M empirical feature spaces is $O(MN^3)$ where N denotes the number of training samples. The huge computational complexity makes MEKL to be incapable of dealing with large-scale problems. Secondly, the dimension of each generated empirical feature space is approximate to N which results in severe memory burden. Thirdly, MEKL utilizes the gradient decent optimization approach [42] to iteratively minimize the loss function, which is time consuming to learn an appropriate classifier. Finally, most MEKL can only deal with binary-class problems making them have to employ some decomposition strategies [17,20], such as one-against-one or one-against-all, to decompose a multi-class problem into multiple binary-class sub-problems, which also leads to the consequence of consuming more time to learn a final classifier. The above problems seriously restrict the application of MEKL on large-scale problems because of the severe computational complexity and memory burden.

In order to overcome these four problems, we propose a fast Multiple Random Empirical Kernel Learning Machine (MREKLM). To solve the first and second problems, MREKLM employs Random Empirical Kernel Mapping (REKM) based on Random Projections (RP) [5] technique, to empirically map the samples into multiple low-dimensional empirical feature spaces with lower computational complexity. RP technique aims to project data from high-dimensional space into a low-dimensional one via a random matrix [1]. Based on RP technique, REKM can efficiently construct M low-dimensional empirical feature spaces with $O(MP^3)$ computational complexity where $P(\ll N)$ represents the number of randomly selected samples to construct a specific empirical feature space. Moreover, to settle the third and fourth problems, MREKLM adopts an analytical approach to optimize the learner. Firstly, MREKLM adopts a weighted combination of empirical feature spaces. Then, it optimizes the augmented weight vector of the classifier by adopting the minimum norm least-squares [26] solution with Moore–Penrose generalized inverse [7], i.e. MREKLM learns the final classifier with an analytical optimization approach which directly deals with a multi-class problem via an analytical style. The whole computational complexity of MREKLM is $O(M^3P^3)$. To validate the effectiveness and efficiency of MREKLM, experiments are conducted on several datasets with tens of thousands of samples. Experimental results indicate that the average training speed of MREKLM for learning a classifier is about 660 times faster than MEKL, and even about 26 times faster than SVM with RBF kernel. Moreover, MREKLM results in a considerable classification performance compared with the implemented algorithms. Therefore, we can conclude that the proposed MREKLM can effectively deal with the problems with much lower computational complexity. According to the Radamacher complexity, MREKLM has similar generalization error bound to MEKL. We further discuss the relationship between the proposed REKM and the Nyström approximation, which demonstrates that REKM and the Nyström approximation are the particular case of the random projection. Moreover, REKM can result in the same effect as the Nyström approximation in dimensionality reduction.

It should be stated that we have proposed an alternative MEKL classifier based on RP technique, named MEKLRP [43], which firstly randomly selects a subset from the training set through RP strategy and adopts the selected subset to construct multiple empirical feature spaces. Then, MEKLRP employs a Modification of Ho–Kashyap algorithm with Squared approximation of the misclassification errors (MHKS) [30] as the base classifier in each empirical feature space. After that, MEKLRP utilizes a regularization term to combine these sub-classifiers. That is, MEKLRP iteratively optimizes the objective function to learn the classifier. Different from MEKLRP, the proposed MREKLM firstly selects multiple random subsets, and utilizes each subset to generate a specific empirical feature space. Then, MREKLM optimizes the classifier via an analytical optimization method. Although both MEKLRP and MREKLM adopt the random subset to generate empirical feature spaces, MEKLRP adopts one random subset, i.e. the construction samples for all empirical feature spaces are the same. While in MREKLM, the selected samples for each empirical feature space are totaly randomly selected, i.e. different empirical feature spaces are constructed by different random subsets. Moreover, in the classifier learning phase, MREKLM and MEKLRP are totally different in both learning objective function and the corresponding optimization methods. Thus, it is evident that the proposal in this paper is different from our previously proposed MEKLRP.

The contributions of this paper are highlighted as follows:

- This paper proposes a fast MEKL algorithm named MREKLM which is of much lower computational complexity. MREKLM learns the classifier via an analytical optimization method and deals with multi-class problem via a direct approach.
- This paper introduces an efficient EKM named REKM which empirically maps the samples into low-dimensional feature spaces with much lower computational complexity. Moreover, the discriminant information of samples can be approximately preserved in the low-dimensional feature spaces, which guarantees the classification performance of the learned classifier.
- This paper extends the capability of MEKL in dealing with large-scale problems. With the random empirical kernel mapping, the computation cost of the proposed method is cubic in MP, where M and P denote the number of feature spaces and the number of randomly selected training samples, respectively.

The rest of this paper is organized as follows: Section 2 presents a brief description on the related work of the proposed MREKLM. Then, Section 3 gives a detailed illustration on the proposed MREKLM. After that, experimental results are discussed and reported in Section 4. Finally, the concluding remarks are presented in Section 5.

2. Related work

This section presents the related work of the proposed MREKLM. We first demonstrate how to calculate the empirical kernel mapping over the whole training samples. Then, we present a brief description on recent researches on random projections technique.