Neurocomputing 149 (2015) 253–264

Contents lists available at ScienceDirect

Neurocomputing

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

# Multiple kernel extreme learning machine

Xinwang Liu [a,*], Lei Wang [b], Guang-Bin Huang [c], Jian Zhang [d], Jianping Yin [a]

[a] School of Computer Science, National University of Defense Technology, Changsha 410073, China
[b] School of Computer Science and Software Engineering, University of Wollongong, NSW 2522, Australia
[c] School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore
[d] Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia

ARTICLE INFO

*Article history:*
Received 2 August 2013
Received in revised form
18 September 2013
Accepted 20 September 2013
Available online 6 September 2014

*Keywords:*
Extreme learning machine
Multiple kernel learning
Support vector machines

ABSTRACT

Extreme learning machine (ELM) has been an important research topic over the last decade due to its high efficiency, easy-implementation, unification of classification and regression, and unification of binary and multi-class learning tasks. Though integrating these advantages, existing ELM algorithms pay little attention to optimizing the choice of kernels, which is indeed crucial to the performance of ELM in applications. More importantly, there is the lack of a general framework for ELM to integrate multiple heterogeneous data sources for classification. In this paper, we propose a general learning framework, termed multiple kernel extreme learning machines (MK-ELM), to address the above two issues. In the proposed MK-ELM, the optimal kernel combination weights and the structural parameters of ELM are jointly optimized. Following recent research on support vector machine (SVM) based MKL algorithms, we first design a sparse MK-ELM algorithm by imposing an $\ell_1$-norm constraint on the kernel combination weights, and then extend it to a non-sparse scenario by substituting the $\ell_1$-norm constraint with an $\ell_p$-norm ($p>1$) constraint. After that, a radius-incorporated MK-ELM algorithm which incorporates the radius of the minimum enclosing ball (MEB) is introduced. Three efficient optimization algorithms are proposed to solve the corresponding kernel learning problems. Comprehensive experiments have been conducted on Protein, Oxford Flower17, Caltech101 and Alzheimer's disease data sets to evaluate the performance of the proposed algorithms in terms of classification accuracy and computational efficiency. As the experimental results indicate, our proposed algorithms can achieve comparable or even better classification performance than state-of-the-art MKL algorithms, while incurring much less computational cost.



## 1. Introduction

Extreme learning machine (ELM) was first designed for single hidden layer feedforward neural networks [1–3] and then extended to generalized single hidden layer feedforward networks (SLFN) which did not necessarily resemble neurons [4,5]. Different from traditional neural SLFN learning algorithms, ELM aims to minimize both training error and the norm of output weights [3,6]. Due to its (1) *high efficiency*, (2) *easy-implementation*, (3) *unification of classification and regression* and (4) *unification of binary and multi-class classification* [6], ELM has been an active research topic over the past a few years [3,6–12]. In addition, the ELM has also been successfully applied to many applications such as imbalance learning [13], missing data learning [14] and activity recognition [15], to name just a few. More recent advances in ELM can be found in [10–12].

Although researchers have made great progress from both a theoretical and a practical point of view, ELM has still not well considered the following two issues. The first one is how to choose an optimal kernel for a specific application when the kernel trick is applied to ELM such as in previous work [6,16–18]. The other one is how to handle information fusion in ELM when multiple heterogenous data sources are available. In this paper, we propose a general framework by borrowing the idea of multiple kernel learning (MKL) to handle the above two issues. We call our framework a multiple kernel extreme learning machine (MK-ELM). In the MK-ELM, the optimal kernel is assumed to be a linear combination of a group of base kernels, and the base kernel combination weights and structural parameters of ELM are jointly optimized in the learning process. Though sharing the same assumption that the optimal kernel is a linear combination of base kernels, the proposed MK-ELM and the widely studied SVM based MKL algorithms have important differences. (1) For the proposed MK-ELM, the binary and multi-class classification problems are unified into one common formula. In contrast, the one-against-one (OAO) and one-against-all (OAA) strategies are usually

* Corresponding author.
*E-mail address:* 1022xinwang.liu@gmail.com (X. Liu).

http://dx.doi.org/10.1016/j.neucom.2013.09.072
0925-2312/

adopted in SVM based MKL algorithms [19,20] to handle the multi-class classification problems. (2) The optimization problem for MK-ELM is much simpler than the one used in SVM based MKL algorithms. The structural parameter of MK-ELM can be analytically obtained by a matrix inverse operation, while a constrained quadratic programming (QP) solver is required to solve the optimization problems of SVM based MKL algorithms.

In the literature, there are mainly three research directions for existing SVM based MKL algorithms, including sparse MKL algorithms [19,21–23], non-sparse MKL algorithms [20,24] and the recent radius-incorporated MKL variants [25–27]. In order to conduct a comprehensive comparison with SVM based MKL algorithms, we also design sparse, non-sparse and radius-incorporated MK-ELM algorithms in this paper. Specifically, the contributions of this paper are highlighted as follows:

1. A sparse MK-ELM algorithm is first developed, where an $\ell_1$-norm constraint is imposed on the base kernel combination weights.
2. A non-sparse variant is proposed by substituting the $\ell_1$-norm constraint with an $\ell_p$-norm constraint, where $p>1$.
3. Another radius-incorporated MK-ELM is then proposed by integrating the radius of minimum enclosing ball (MEB) [28,29] into the objective function of MK-ELM.
4. Comprehensive experiments have been conducted to compare the proposed MK-ELM variants with existing state-of-the-art MKL algorithms, including multiple kernel SVM (MK-SVM) [19], multiple kernel least square SVM (MK-LSSVM) [30], multiple kernel fisher discriminative analysis (MK-FDA) [23], and their sparse and non-sparse variants. The experimental results demonstrate that the proposed MK-ELM variants achieve statistically comparable or better classification performance while requiring less training time.

The rest of this paper is organized as follows. We review the extreme learning machine and multiple kernel learning in Section 2. In Section 3, we first present the formulation of the sparse MK-ELM, extend it to a non-sparse case and then propose a radius-incorporated variant. Three efficient algorithms are given to solve the resulting optimization problems. Extensive experimental comparison is conducted in Sections 4 and 5 draws our conclusion.

## 2. Related work

In this section, we give a brief review of extreme learning machine and multiple kernel learning. Though ELM unifies classification and regression tasks, we only focus on classification in the following parts.

### 2.1. Extreme learning machine

According to the ELM theory [6,10], ELM aims to simultaneously minimize the training errors and the norm of output weights. This objective function, for both binary and multi-class classification tasks, can be expressed as follows:

$$\min_{\boldsymbol{\beta},\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|_{\mathbf{F}}^2+\frac{C}{2}\sum_{i=1}^{n}\|\boldsymbol{\xi}_{\cdot i}\|^2 \quad \text{s.t. } \boldsymbol{\beta}^\top\boldsymbol{\phi}(\mathbf{x}_i)=\mathbf{y}_i-\boldsymbol{\xi}_{\cdot i}, \quad \forall i, \tag{1}$$

where $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$ is a training set, $\boldsymbol{\phi}(\mathbf{x}_i)(i=1,\ldots,n)$ is the hidden-layer output (feature mapping) corresponding to $\mathbf{x}_i$, $\boldsymbol{\beta}\in\mathbb{R}^{|\boldsymbol{\phi}(\cdot)|\times T}$ is the output weights, $\boldsymbol{\xi}\in\mathbb{R}^{T\times n}$ is the training error matrix on training data, $\boldsymbol{\xi}_{\cdot i}=[\xi_{1i},\xi_{2i},\ldots,\xi_{Ti}]^\top(1\le i\le n)$ is the $i$th column of $\boldsymbol{\xi}$, $\mathbf{y}_i=[0,\ldots,0,\underset{t}{1},0,\ldots,0]^\top\in\{0,1\}^T$ if $\mathbf{x}_i$ belongs to the $t$th ($1\le t\le T$) class, $n$ and $T$ are the number of training samples and classes, and $C$ is a regularization parameter which trades off the norm of output weights and training errors. $\|\cdot\|_{\mathbf{F}}$ is the Frobenius norm.

The optimization problem in Eq. (1) can be efficiently solved. According to [6], the optimal $\boldsymbol{\beta}^\star$ which minimizes Eq. (1) can be analytically obtained as

$$\boldsymbol{\beta}^\star=\boldsymbol{\Phi}^\top\left(\frac{\mathbf{I}}{C}+\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\right)^{-1}\mathbf{Y}^\top, \tag{2}$$

where $\boldsymbol{\Phi}=[\boldsymbol{\phi}(\mathbf{x}_1),\ldots,\boldsymbol{\phi}(\mathbf{x}_n)]^\top\in\mathbb{R}^{n\times|\boldsymbol{\phi}(\cdot)|}$, $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]\in\mathbb{R}^{T\times n}$ and $\mathbf{I}$ is an identity matrix.

As can be seen from the above, both the binary and multi-class classification tasks in ELM can be handled via an unified formula Eq. (1). Moreover, Eq. (1) can be analytically solved by a matrix inverse operation, while a constrained quadratic programming problem is required in SVM. This makes the ELM easy and efficient to implement due to the fact that solving a matrix inverse problem is usually much more computationally efficient than solving the same-size constrained QP problem. In addition, it is worth mentioning that though both ELM and least square SVM (LSSVM) [31] share the same objective function as far as the optimization is concerned, there is no bias term deployed in ELM, as in Eq. (1). Such a subtle difference makes ELM to have milder optimization constraint than LSSVM. These advantages help ELM to achieve better classification performance while incurring less computational cost, as demonstrated by the experimental results in [6].

After obtaining the optimal $\boldsymbol{\beta}^\star$, the decision score of the ELM on test point $\mathbf{x}$ is determined by

$$f(\mathbf{x})=\boldsymbol{\beta}^{\star\top}\boldsymbol{\phi}(\mathbf{x}), \tag{3}$$

and the index corresponding to the highest value of $f(\mathbf{x})\in\mathbb{R}^T$ is considered as the label of $\mathbf{x}$.

### 2.2. Multiple kernel learning

It is well known that the choice of kernels is crucial for kernel-based algorithms [32]. Much effort has been devoted to tuning an optimal kernel for a specific application [19,33,27]. MKL provides an elegant way to handle such an issue by optimizing a data-dependent kernel. In MKL, the optimal kernel is assumed to be a linear combination of a group of base kernels, and the optimal combination coefficients and the structural parameters of classifiers are jointly learned by maximizing the margin [19,30], class separability criterion [24,23], etc. Specifically, MKL takes the form of

$$\kappa(\cdot,\cdot;\boldsymbol{\gamma})=\sum_{p=1}^{m}\gamma_p\kappa_p(\cdot,\cdot), \tag{4}$$

where $\{\kappa_p(\cdot,\cdot)\}_{p=1}^m$ are $m$ pre-defined base kernels, and $\{\gamma_p\}_{p=1}^m$ are the base kernel combination coefficients. Eq. (4) can be equivalently rewritten as

$$\boldsymbol{\phi}(\cdot;\boldsymbol{\gamma})=[\sqrt{\gamma_1}\boldsymbol{\phi}_1(\cdot),\sqrt{\gamma_2}\boldsymbol{\phi}_2(\cdot),\ldots,\sqrt{\gamma_m}\boldsymbol{\phi}_m(\cdot)], \tag{5}$$

where $\boldsymbol{\phi}(\cdot;\boldsymbol{\gamma})$ and $\{\boldsymbol{\phi}_p(\cdot)\}_{p=1}^m$ are the feature mappings corresponding to kernels $\kappa(\cdot,\cdot;\boldsymbol{\gamma})$ and $\{\kappa_p(\cdot,\cdot)\}_{p=1}^m$, respectively.

Usually, a constraint is imposed on the kernel combination weights $\boldsymbol{\gamma}$ to make the optimization problems bounded and the combined kernel be positive semi-definite (PSD). One common example is imposing an $\ell_q(q=1)$ norm and non-negative constraint on the kernel combination weights. Such constraint will induce sparse kernel combination, as shown in [19,21–23]. Another one is imposing an $\ell_q(q>1)$ norm and non-negative constraint. Unlike the previous one, this constraint will bring forth non-sparse kernel combination [20,24]. In the following section, we will design sparse and non-sparse multiple kernel learning algorithms for ELM by varying $q$ from one to any positive number larger than one.