Computing and Informatics, Vol. 36, 2017, 618–636, doi: 10.4149/cai_2017_3_618

COUPLED MULTIPLE KERNEL LEARNING FOR SUPERVISED CLASSIFICATION

En Zhu, Qiang Liu, Jianping Yin

College of Computer
National University of Defense Technology
Changsha, Hunan, China 41073
e-mail: {enzhu, qiangliu06, jpyin}@nudt.edu.cn

Abstract. Multiple kernel learning (MKL) has recently received significant attention due to the fact that it is able to automatically fuse information embedded in multiple base kernels and then find a new kernel for classification or regression. In this paper, we propose a coupled multiple kernel learning method for supervised classification (CMKL-C), which comprehensively involves the intra-coupling within each kernel, inter-coupling among different kernels and coupling between target labels and real ones in MKL. Specifically, the intra-coupling controls the class distribution in a kernel space, the inter-coupling captures the co-information of base kernel matrices, and the last type of coupling determines whether the new learned kernel can make a correct decision. Furthermore, we deduce the analytical solutions to solve the CMKL-C optimization problem for highly efficient learning. Experimental results over eight UCI data sets and three bioinformatics data sets demonstrate the superior performance of CMKL-C in terms of the classification accuracy.

Keywords: Multiple kernel learning, non-IIDness, coupled kernels, supervised classification

1 INTRODUCTION

In the big data era, it is vital to extract and aggregate diverse information that is embedded in a huge volume of data with different characteristics in order to deduce some high-level knowledge. For example, the sentiment of a person can be determined via jointly analyses of his/her published texts, pictures, and videos in

social media. Among a variety of machine learning tools, multiple kernel learning (MKL) is good at knowledge discovery from heterogeneous big data. The MKL is a novel machine learning paradigm that enables classifiers or regression functions to utilize different kinds of information embedded in multiple base kernels. Given a set of base kernels, the goal of MKL is to construct a new kernel, which is more suitable to address the problem at hand, through learning the optimal combination coefficients of multiple base kernels. Due to the high classification performance of MKL in a variety of application scenarios, such as image classification [1], bioinformatics [2] and video event detection [3], numerous researchers have been devoting themselves to not only the theoretical research but also the application of the MKL to further improve its overall performance. However, existing MKL methods generally made an assumption of IIDness, i.e., independent and identically distributed. The assumption means that all kernels, attributes and their values are independent and follow identical distributions. However, the IIDness assumption ignores lots of coupling relationships [14] embedded in multiple kernels, including the intra-coupling within each base kernel, the inter-coupling between base kernels and the coupling between learning objectives and base kernels. Therefore, introducing the above three coupling relationships into the MKL method is beneficial for capturing complex relationships among data and can further improve the learning performance.

To overcome the weakness of conventional learning methods that were based on the IIDness assumption, some non-IIDness learning methods were proposed for advanced big data analytics, for example the non-IIDness learning in behavioral and social data [15]. As stated in [15], the main task of non-IIDness learning was to learn complex couplings and heterogeneity, which were two significant features of non-IID data. Therefore, we argue that the overall performance of the MKL method for classification can be further improved by releasing the IIDness assumption and jointly considering the intra-coupling within each base kernel, the inter-coupling between base kernels and the coupling between learning objectives and base kernels. Hence, we propose a coupled multiple kernel learning method for classification (CMKL-C) in this paper to meet the above mentioned requirement. Specifically, we present a new CMKL-C objective function that jointly considers the above three couplings to learn a more powerful kernel. Then, we deduce the analytical solution of optimizing the CMKL-C objective function, which guarantees the high learning efficiency of CMKL-C. The coupling relationships include the following three aspects:

The intra-coupling within each kernel: Such coupling embodies the sample distribution. Specifically, it measures both the sample within-class variance and the sample-class center distance at the same time in order to reflect whether a kernel has the separable ability. For kernel based classification, the more separable ability a kernel has, the higher performance it can achieve. Thus, CMKL-C adjusts the weights of multiple base kernels based on their intra-couplings so that the weights of those kernels with more separable ability can be assigned with larger values.

The inter-coupling among different kernels: Such coupling reflects the co-information shared among different base kernels. Specifically, it utilizes a kernel alignment method to calculate the similarities among base kernels. Then, it treats the nearing density of each kernel in the multiple kernel space as the inter-coupling metric according to their similarities. A kernel that has a higher inter-coupling value shares more common information with other kernels. Since the MKL aims to fuse heterogeneous information involved in these base kernels, it should balance the weights of base kernels to make sure that the common information is not unique to any kernel. Therefore, CMKL-C regulates base kernels via combining coefficients with other kernels.

The coupling between target labels and real ones: Such coupling represents the similarities between the target labels of sample data determined by MKL and their real labels. It is worth noticing that the goal of MKL is to learn a new kernel that can predict the right labels of all sample data. Therefore, CMKL-C utilizes such coupling to optimize the learning performance of MKL. As far as we know, the proposed CMKL-C is the first non-IIDness learning method that couples multiple kernels for supervised classification. To validate the superior performance of the proposed CMKL-C method, eight UCI classification data sets and three bioinformatics benchmark data sets are used to test the performance of the proposed CMKL-C method. Experimental results demonstrate that CMKL-C significantly outperforms the state-of-the-art MKL methods in terms of the classification accuracy.

The rest of this paper is organized as follows: Section 2 briefly reviews the state-of-the-art work regarding the MKL and the non-IIDness learning. Section 3 introduces the formal definitions with respect to three coupling relationships in MKL. Then, Section 4 shows the proposed coupled multiple kernel learning method for classification (CMKL-C). After that, Section 5 evaluates the performance of CMKL-C over eight UCI classification data sets and three bioinformatics benchmark data sets. Finally, Section 6 gives some conclusion remarks and presents our future work.

2 RELATED WORK

2.1 Multiple Kernel Learning

In recent years, multiple kernel learning (MKL) has made significant progress, such as high classification accuracy, automatic kernel parameter setting, and multi-source information fusion, etc. Therefore, an increasing number of researchers focus on not only the theoretical research but also the application of MKL. For example, Harchaoui et al. applied MKL into image classification [1], Liu et al. introduced MKL into Alzheimer's disease prediction [2], and Vahdat et al. adopted MKL into video event detection [3], etc. Typically, the MKL work can be categorized into two groups as per its learning approaches:

1. One-stage approach. The one-stage MKL method jointly learns the combination coefficients of multiple base kernels and the classifier parameters by seeking a jointly optimization objective function. The one-stage MKL method was firstly investigated in [4]. Then, numerous one-stage MKL methods were proposed to improve its classification capacity and its efficiency further, e.g. [5, 6, 7, 8, 9, 10].
2. Two-stage approach. The two-stage MKL method [11, 12, 13] first finds a suitable combination strategy of multiple base kernels, and then it uses the combination strategy to construct a new kernel that serves as the resulting classifier/regression function.

Although the state-of-the-art MKL methods achieved considerable performance in different application scenarios, they all made the above mentioned IIDness assumption. Therefore, the conventional MKL methods ignored the intra-coupling within each base kernel, the inter-coupling between base kernels and the coupling between learning objectives and base kernels.

2.2 Non-IIDness Learning

Non-IIDness is a brand new learning paradigm, which was proposed for high performance big data analytics [15] by capturing the intrinsic data characteristics and complexities. The authors in [15] analyzed the characteristics of non-IID data and pointed out that coupling and heterogeneity were two significant characteristics of non-IID data, and the main task of non-IIDness learning was to learn complex couplings and heterogeneity, which laid a solid theoretical framework for the non-IIDness learning study. Meanwhile, the author discussed problems that may be caused by adopting classical learning approaches based on the IIDness assumption to learn non-IID data, revealing the advantages of the non-IIDness learning research. The formal definition of coupled behaviors in non-IID data was first proposed on the basis of coupled hidden Markov model [16]. After that, a coupling learning method with more complex interactions was proposed in [14]. Different coupling measuring methods had been proposed for different kinds of data in [17, 18]. In [17], a coupled nominal similarity was introduced into unsupervised learning, while the similarity metric proposed in [18] captured the coupled attributes of numerical data via a Taylor expansion likely method, which delicately combined the intra-coupling in each attribute and the inter-coupling among different attributes. At a higher level, coupled ensemble clustering in [19] used coupling relationships among both base clustering results and objects to merge different clustering results.

3 PRELIMINARIES

3.1 Intra-Coupling within Each Kernel

In kernel-based learning, a kernel is generally used to map features into a high dimensional space such that different classes can be separated linearly. In other words,

for a certain set of samples, the samples within a class will locate in a similar distribution while the samples from different classes will locate in different distributions after mapping by a well-defined kernel. Such distribution information contained in each kernel matrix reflects the intra-coupling of the kernel, which is the coupling relationship among samples' classes. It is clear that the intra-coupling of each kernel reflects the goodness of the kernel.

We measure the intra-coupling of each kernel by considering the distribution of samples in a kernel space. On one hand, we emphasize the concentration of samples within the same class after kernel mapping. For a well-defined kernel, the samples in the same class should be as concentrated as possible. Here, we introduce the variance to measure such a characteristic. On the other hand, we are also concerned about the separation of samples from different classes after kernel mapping. A suitable kernel can always guarantee large distance between two samples from different classes. Therefore, we use the distance between two classes to reflect the dispersion in CMKL-C. Overall, the intra-coupling of each kernel is measured by combining the variance of samples within a class with the distance between two classes, which can also be seen as the intra-coupling and the inter-coupling of classes, respectively.

In this paper, we follow the measurement proposed in [20], in which the distribution mentioned above is used to evaluate the goodness of the kernel matrix. Firstly, we consider how to measure the variance of samples within a class in a kernel space for binary classification. Typically, the standard deviation is used to measure the variance of a distribution. However, the variance in the direction of separating hyperplane does not affect the classification performance. A good measurement that is suitable for classification task is the standard deviation in the direction of between-class centers. We denote this standard deviation as *std* in this paper. In each class, we can also calculate the standard deviation of data distribution in this direction and denote them as std_+ and std_- for the first and second class of data respectively. Hence, given a kernel $k(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$, where $\phi(\cdot)$ is a feature mapping function, we can calculate the data center of two classes ϕ_+ and ϕ_- via $\phi_+ = \sum_{i=1}^{n_+} \phi(x_i)/n_+$ and $\phi_- = \sum_{i=n_++1}^{n} \phi(x_i)/n_-$, respectively. Since the unit vector in the direction of between-class centers is $\mathbf{u} = \frac{\phi_- - \phi_+}{\|\phi_- - \phi +\|}$, the std_+ and std_- can be calculated as

$$
\begin{aligned}
std_+ &= \sqrt{\frac{\sum_{i=1}^{n_+} \langle \phi(x_i) - \phi_+, \mathbf{u} \rangle^2}{n_+ - 1}} = \sqrt{\frac{\sum_{i=1}^{n_+} \langle \phi(x_i) - \phi_+, \phi_- - \phi_+ \rangle^2}{(n_+ - 1)(\phi_- - \phi_+)^2}} \\
&= \sqrt{\frac{\sum_{i=1}^{n_+} (\phi(x_i)\phi_- + \phi_+^2 - \phi(x_i)\phi_+ - \phi_+\phi_-)^2}{(n_+ - 1)(\phi_- - \phi_+)^2}}, \qquad (1)
\end{aligned}
$$

and

$$std_- = \sqrt{\frac{\sum_{i=n_++1}^{n}\langle\phi(x_i)-\phi_-,\mathbf{u}\rangle^2}{n_- - 1}} = \sqrt{\frac{\sum_{i=n_++1}^{n}\langle\phi(x_i)-\phi_-,\phi_- - \phi_+\rangle^2}{(n_- - 1)(\phi_- - \phi_+)^2}}$$

$$= \sqrt{\frac{\sum_{i=n_++1}^{n}(\phi(x_i)\phi_+ + \phi_-^2 - \phi(x_i)\phi_- - \phi_-\phi_+)^2}{(n_- - 1)(\phi_- - \phi_+)^2}}. \tag{2}$$

For the convenience of description, we define some auxiliary variables as follows,

$$\begin{aligned}
a_i &= \phi(x_i)\phi_+ = \frac{\sum_{j=1}^{n_+}\phi(x_i)\phi(x_j)}{n_+} = \frac{\sum_{j=1}^{n_+}k_{ij}}{n_+},\\
&\quad (i = 1,\ldots,n_+),\\
b_i &= \phi(x_i)\phi_- = \frac{\sum_{j=n_++1}^{n}\phi(x_i)\phi(x_j)}{n_-} = \frac{\sum_{j=n_++1}^{n}k_{ij}}{n_-},\\
&\quad (i = 1,\ldots,n_+),\\
c_i &= \phi(x_i)\phi_+ = \frac{\sum_{j=1}^{n_+}\phi(x_i)\phi(x_j)}{n_+} = \frac{\sum_{j=1}^{n_+}k_{ij}}{n_+},\\
&\quad (i = n_+ + 1,\ldots,n),\\
d_i &= \phi(x_i)\phi_- = \frac{\sum_{j=n_++1}^{n}\phi(x_i)\phi(x_j)}{n_-} = \frac{\sum_{j=n_++1}^{n}k_{ij}}{n_-},\\
&\quad (i = n_+ + 1,\ldots,n),\\
A &= \phi_+\phi_+ = \frac{\sum_{i=1}^{n_+}a_i}{n_+},\\
B &= \phi_+\phi_- = \frac{\sum_{i=1}^{n_+}b_i}{n_+},\\
C &= \phi_-\phi_+ = \frac{\sum_{i=n_++1}^{n}c_i}{n_-},\\
D &= \phi_-\phi_- = \frac{\sum_{i=n_++1}^{n}d_i}{n_-}.
\end{aligned} \tag{3}$$

After that, the total standard deviation in the direction of between-class centers can be written as

$$\begin{aligned}
std &= std_+ + std_-\\
&= \sqrt{\frac{\sum_{i=1}^{n_+}(b_i - a_i + A - B)^2}{(n_+ - 1)(A + D - B - C)}} + \sqrt{\frac{\sum_{i=n_++1}^{n}(c_i - d_i + D - C)^2}{(n_- - 1)(A + D - B - C)}}.
\end{aligned} \tag{4}$$

Regarding the distance between binary classes, we select the distance between centers of two classes as the inter-class distance. This distance is suitable for reflecting the inter-class distance for classification task when the standard deviation in the direction of between-class centers is used to measure the variance of class. The inter-class distance can be calculated as follows,

$$dist = \|\phi_- - \phi_+\|. \tag{5}$$

As we discussed above, a well-defined kernel should map samples into a space in which the variance of samples within the same class is small while the inter-class distance is large. Thus, we measure the intra-coupling of kernel for the binary classification case by calculating the ratio of the total standard deviation in the direction of between-class centers of the class-pair to the distance between the centers of two classes:

$$intra = \frac{std}{dist}. \tag{6}$$

Regarding multiple classification case, we use (6) to calculate such ratio for all class-pairs and denote the intra-coupling as the mean value of them. Assuming there are m base kernels, we can calculate intra-coupling for each of them and construct a m-dimension intra-coupling vector:

$$\mathbf{intra} = \{intra_1, \ldots, intra_m\}^\top. \tag{7}$$

Since the intra-coupling of a kernel reflects the goodness of the kernel, we can use it to guide MKL design. If a base kernel matrix has a small intra-coupling value, we regard it as an important kernel in learning, because it either corresponds to a separable space or owns important information for the classification task. Therefore, one of the optimization objectives is to minimize the intra-coupling of multiple kernels.

3.2 Inter-Coupling Among Different Kernels

Currently, most of MKL approaches treat base kernels independently when learning their combinations. However, inter-coupling among base kernels is important to be considered in the learning approach. In the following part, we will firstly discuss the inter-coupling among base kernels and then describe how to measure the inter-coupling and use it in MKL.

We consider three types of multiple kernel matrix generation methods:

1. using different kernel functions to map the same channel of samples;
2. using one kernel function to map different channels of samples;
3. using different kernel functions to map different channels of samples.

No matter what the generation method is, we can see that these kernel matrices all reflect a profile of the same samples using the same or different measurements. Thus, the same samples inevitability originate the coupling relationship among base

kernels. These inter-couplings represent that base kernel matrices share parts of the same sample information. The more information the kernel matrices share, the larger degree of inter-coupling they have.

To measure the inter-coupling among base kernels, we first calculate the co-information of kernels through kernel alignment [11], which has great theoretical properties for measuring kernel matrix similarity. Given two kernel matrices $\mathbf{K}_1$ and $\mathbf{K}_2$, which both map the sample set $\mathbf{S}$, the alignment between them can be calculated as follows,

$$\hat{A}(\mathbf{S}, \mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle_F \langle \mathbf{K}_2, \mathbf{K}_2 \rangle_F}} \tag{8}$$

where $\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F = \sum_{i,j=1}^{n} \mathbf{K}_1(x_i, x_j)\mathbf{K}_2(x_i, x_j)$ and n is the number of samples. Assuming that there are m base kernel matrices $\{\mathbf{K}_1, \ldots, \mathbf{K}_m\}$, we define the inter-coupling of $\mathbf{K_i}$ with others as follows,

$$inter_i = \sum_{j=1}^{m} e^{\left(-\frac{1-\hat{A}(\mathbf{S}, \mathbf{K}_i, \mathbf{K}_j)}{\sigma}\right)^2} \tag{9}$$

where e is the mathematical constant and σ is a parameter that control the nearing distance. Equation (9) uses Gaussian function to calculate the distance between each base kernel in kernel spaces. As a result, for a specific kernel $\mathbf{K_i}$, other kernels near to it will induce a large value of $inter_i$. Moreover, the value of $inter_i$ significantly decreases along with the increase of the distance between $\mathbf{K_i}$ and other kernels. In this case, the kernel with more nearing kernels will have larger inter-coupling value. Thus, the inter-coupling measures the density of the kernel in multiple kernel spaces.

Then we can get the inter-coupling vector of base kernels:

$$\mathbf{inter} = (inter_1, \ldots, inter_m)^\top. \tag{10}$$

MKL expects to confuse all information that contented by base kernels. Typically, it linearly combines base kernel matrices to capture all information. However, some information that is shared by kernel matrices may be aggregated using linear combination. Consequently, this information may include the unique information that is contained in few kernels. Hence, a consideration solution is to regulate the kernel weights by considering their inter-coupling. Specifically, we should ensure that the weights of those kernels with low inter-coupling are large while the weights of the kernels with high inter-coupling are small. By this method, the MKL approach can simultaneously capture the common information and preserve the unique information, resulting in a more powerful kernel. In CMKL-C, the inter-coupling of base kernel matrices should be also minimized in the objective function.

3.3 Coupling Between Target Labels and Real Ones

Finally, we consider the coupling between the new kernel, which is learned through MKL and real labels. Obviously, a label represents the category of a group of samples. Different categories will have different labels. Moreover, the kernel matrix also implies the category information of samples. Specifically, the kernel matrix entry with a small value means that two samples come from the same category. On the contrary, a large value of the kernel matrix entry suggests that the corresponding samples come from different categories. Thus, the real labels reflect the real category information while the target labels determine the categories that are classified by the new learned kernel. Hence, if the coupling between target labels and real ones is stronger, the classification result will be better. In this paper, we consider this coupling in multiple kernel learning.

For the binary classification case, the label gives the binary information of categories, i.e. either 1 or -1. Accordingly, the kernel matrix involves two classes in kernel space and represents it as its entry values. The objective is to make the learned kernel determine the same category information with respect to the samples as their real labels. In other words, if two samples are from the same category, the entry values with respect to them should be 1, which means they are extremely similar in the kernel space. Otherwise, we hope the values be -1. Thus, we define the coupling between target labels and real ones as

$$between_{i,j} = 1 - |k_{i,j} - t_{i,j}| \tag{11}$$

where $k_{i,j}$ is the entry value of the new kernel with respect to sample i and sample j, $t_{i,j}$ corresponds to the label-pair as follows,

$$t_{i,j} = \begin{cases} 1 & label_i = label_j, \\ -1 & label_i \neq label_j. \end{cases} \tag{12}$$

Therefore, one of the optimization objectives is to maximize the coupling between target labels and real ones.

For the multiple classification case, the label can be seen as a nominal value that reflects the information of categories. Labels can give the information that whether samples are from the same category but cannot tell us that how similar two categories are. Fortunately, kernel matrix can display this similarity directly. Therefore, for a pair of samples from the same category, the coupling between target labels and real ones can be defined similar to the case of binary classification:

$$\begin{aligned} &between_{i,j} = 1 - |k_{i,j} - 1|, \\ &\text{s.t. } label_i = label_j. \end{aligned}$$

For a pair of samples from different categories, we assume that their similarity should be extremely small. Thus, the smaller a kernel entry value is, the larger the corresponding coupling value will be. Hence, we define the coupling as follows:

$$\begin{aligned} &between_{i,j} = 1 - |k_{i,j}|, \\ &\text{s.t. } label_i \neq label_j. \end{aligned}$$

4 THE PROPOSED COUPLED MULTIPLE KERNEL LEARNING METHOD FOR CLASSIFICATION

Based on the three coupling relationships mentioned above, we construct an objective function to minimize the former two types of coupling but maximize the last type of coupling. The objective function is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{\mu}} \quad & \tfrac{1}{2}\|\boldsymbol{\mu}\|^2 + C_1 \mathbf{Intra}^\top \boldsymbol{\mu} + C_2 \mathbf{Inter}^\top \boldsymbol{\mu} \\ & + \tfrac{1}{2} C_3 \textstyle\sum_{i=1}^{n} \xi_i^2, \\ \text{s.t.} \quad & \mathbf{z}_i^\top \boldsymbol{\mu} = t_i - \xi_i, \quad i = \{1, 2, \ldots, n\}. \end{aligned} \tag{13}$$

In Equation (13), $\boldsymbol{\mu}$, which needs to be learned, is the combination coefficient of base kernels. **Intra** is the intra-coupling vector and **Inter** is the inter-coupling vector. Meanwhile, $\mathbf{z}_i = (k_{ik}^{(1)}, k_{ik}^{(2)}, \ldots, k_{ik}^{(m)})^\top$ is a vector containing values from the same entry of different base kernels. Thus, $\mathbf{z}_i\boldsymbol{\mu}$ equals to the entry value of the new kernel. We also denote the coupling between the target label and the real one as ξ. In addition, C_1, C_2 and C_3 are trade-off parameters that control the weights of these three coupling relationships. For the multiple classification case, the objective function is similar to Equation (13) but has some differences in the part of the coupling relationship between target labels and real ones. Accordingly, the objective function is defined as

$$\begin{aligned} \min_{\boldsymbol{\mu}} \quad & \tfrac{1}{2}\|\boldsymbol{\mu}\| + C_1 \mathbf{Intra}^\top \boldsymbol{\mu} + C_2 \mathbf{Inter}^\top \boldsymbol{\mu} \\ & + \tfrac{1}{2} C_3 \left(\textstyle\sum_{i=1}^{same} \xi_i^2 + \sum_{i=same+1}^{n} \gamma_i^2 \right), \\ \text{s.t.} \quad & \mathbf{z}_i^\top \boldsymbol{\mu} = 1 - \xi_i, \quad i = \{1, \ldots, same\}, \\ & \mathbf{z}_i^\top \boldsymbol{\mu} = \gamma_i, \quad i = \{same + 1, \ldots, n\}. \end{aligned} \tag{14}$$

Here, we reorder the kernel entries. Entries for samples from the same category are reordered to the first *same* positions while others are reordered to the $same + 1$ to n positions. In this case, ξ is the coupling between target labels and real ones with respect to the sample pairs from the same category while γ is the coupling of samples pairs from different categories.

Both Equation (13) and Equation (14) have analytical solutions and can be solved efficiently. Regarding Equation (13), solving it is equivalent to solving its dual optimization problem Equation (15) according to the KKT theorem [21].

$$\begin{aligned} L \quad &= \tfrac{1}{2}\|\boldsymbol{\mu}\|^2 + C_1\mathbf{Intra}^\top\boldsymbol{\mu} + C_2\mathbf{Inter}^\top\boldsymbol{\mu} \\ &\quad + \tfrac{1}{2}C_3\textstyle\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i\left(\mathbf{z}_i^\top\boldsymbol{\mu} - t_i + \xi_i\right) \end{aligned} \tag{15}$$

where α_i is the Lagrange multiplier with respect to the i^{th} sample-pair. In this case, the KKT optimality conditions can be written as follows:

$$\frac{\partial L}{\partial \boldsymbol{\mu}} = 0 \rightarrow \boldsymbol{\mu} = -C_1\mathbf{Intra} - C_2\mathbf{Inter} + \sum_{i=1}^{n}\alpha_i\mathbf{z}_i, \tag{16}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_i = C_3\xi_i, \quad i = 1, \ldots, n, \tag{17}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \xi_i = t_i - \mathbf{z}_i^\top\boldsymbol{\mu}, \quad i = 1, \ldots, n. \tag{18}$$

By substituting (18) into (17), we have

$$\begin{aligned} \boldsymbol{\alpha} = C_3\mathbf{t} - C_3\mathbf{Z}\boldsymbol{\mu}: \quad &\alpha_i = C_3 t_i - C_3\mathbf{z}_i^\top\boldsymbol{\mu}, \\ &(i = 1, \ldots, n) \end{aligned} \tag{19}$$

where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^\top$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top$ and $\mathbf{t} = [t_1, t_2, \ldots, t_n]^\top$. Moreover, Equation (16) can be equivalently written as

$$\boldsymbol{\mu} = -C_1\mathbf{Intra} - C_2\mathbf{Inter} + \mathbf{Z}^\top\boldsymbol{\alpha}. \tag{20}$$

By substituting (19) into (20), we have the analytical solution of (13) as follows:

$$\boldsymbol{\mu} = (\mathbf{I} + C_3\mathbf{Z}^\top\mathbf{Z})^{-1}\left(-C_1\mathbf{Intra} - C_2\mathbf{Inter} + C_3\mathbf{Z}^\top\mathbf{t}\right) \tag{21}$$

where $\mathbf{I}$ is a n-dimensional identity matrix. Similarly, solving Equation (14) is equivalent to solving its dual optimization problem Equation (22).

$$\begin{aligned} L = \quad &\tfrac{1}{2}\|\boldsymbol{\mu}\|^2 + C_1\mathbf{Intra}^\top\boldsymbol{\mu} + C_2\mathbf{Inter}^\top\boldsymbol{\mu} + \tfrac{1}{2}C_3\textstyle\sum_{i=1}^{same}\xi_i^2 \\ &+ \tfrac{1}{2}C_3\textstyle\sum_{i=same+1}^{diff}\gamma_i^2 - \sum_{i=1}^{same}\alpha_i\left(\mathbf{z}_i^\top\boldsymbol{\mu} - 1 + \xi_i\right) \\ &- \textstyle\sum_{i=same+1}^{n}\beta_i\left(\mathbf{z}_i^\top\boldsymbol{\mu} - \gamma_i\right) \end{aligned} \tag{22}$$

where α_i and β_i are the Lagrange multiplier with respect to the i^{th} sample pair. In this case, the KKT optimality conditions can be written as follows:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\mu}} = 0 \rightarrow \boldsymbol{\mu} = &- C_1\mathbf{Intra} - C_2\mathbf{Inter} \\ &+ \sum_{i=1}^{same} \alpha_i \mathbf{z}_i + \sum_{i=same+1}^{n} \beta_i \mathbf{z}_i,\end{aligned} \tag{23}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_i = C_3\xi_i, \quad i = 1, \ldots, same, \tag{24}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \xi_i = 1 - \mathbf{z}_i^\top \boldsymbol{\mu}, \quad i = 1, \ldots, same, \tag{25}$$

$$\frac{\partial L}{\partial \gamma_i} = 0 \rightarrow \beta_i = -C_3\gamma_i, \quad i = same + 1, \ldots, n, \tag{26}$$

$$\frac{\partial L}{\partial \beta_i} = 0 \rightarrow \gamma_i = \mathbf{z}_i^\top \boldsymbol{\mu}, \quad i = same + 1, \ldots, n. \tag{27}$$

By substituting (25) into (24), we have

$$\begin{aligned}\boldsymbol{\alpha} = C_3\mathbf{1} - C_3\mathbf{Z}_{same}\boldsymbol{\mu} : \quad &\alpha_i = C_3 - C_3\mathbf{z}_i^\top \boldsymbol{\mu}, \\ &(i = 1, \ldots, same)\end{aligned} \tag{28}$$

where $\mathbf{Z}_{same} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{same}]^\top$ and $\mathbf{1} = [1, 1, \ldots, 1]^\top$ and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{same}]^\top$.

By substituting (27) into (26), we have

$$\begin{aligned}\boldsymbol{\beta} = -C_3\mathbf{Z}_{diff}\boldsymbol{\mu} : \quad &\beta_i = -C_3\mathbf{z}_i^\top \boldsymbol{\mu}, \\ &(i = same + 1, \ldots, n)\end{aligned} \tag{29}$$

where $\mathbf{Z}_{diff} = [\mathbf{z}_{same+1}, \mathbf{z}_{same+2}, \ldots, \mathbf{z}_n]^\top$ and $\boldsymbol{\beta} = [\beta_{same+1}, \beta_{same+2}, \ldots, \beta_n]^\top$. In addition, Equation (23) can be equivalently written as

$$\boldsymbol{\mu} = -C_1\mathbf{Intra} - C_2\mathbf{Inter} + \mathbf{Z}_{same}^\top \boldsymbol{\alpha} + \mathbf{Z}_{diff}^\top \boldsymbol{\beta}. \tag{30}$$

By substituting (28) into (30), we have the analytical solution of (14) as follows:

$$\begin{aligned}\boldsymbol{\mu} = \quad &(\mathbf{I} + C_3\mathbf{Z}_{same}^\top \mathbf{Z}_{same} + C_3\mathbf{Z}_{diff}^\top \mathbf{Z}_{diff})^{-1} \\ &\cdot(-C_1\mathbf{Intra} - C_2\mathbf{Inter} + C_3\mathbf{Z}_{same}^\top \mathbf{1}).\end{aligned} \tag{31}$$

After optimizing the combination coefficient $\boldsymbol{\mu}$, CMKL-C generates a new kernel by combining base kernel matrices according to $\boldsymbol{\mu}$ as follows:

$$\mathbf{K}_{new} = \mu_1\mathbf{K}_1 + \ldots + \mu_m\mathbf{K}_m. \tag{32}$$

Similarly, for each sample-pair (x_i, x_j), its new optimal kernel function can be written as:

$$k_{new}(x_i, x_j) = \mu_1 k_1(x_i, x_j) + \ldots + \mu_m k_m(x_i, x_j). \tag{33}$$

For the classification problem, CMKL-C first considers the intra-coupling of each kernel, inter-coupling of different kernels and the coupling between target labels and real ones. Then, it learns the base kernel combination coefficient according to these three coupling relationships. Finally, it generates a new kernel by combining base kernels according to the above coefficient. The algorithm of CMKL-C is illustrated as Algorithm 1.

Algorithm 1 CMKL-C

1: **Input**: base kernel matrices $\mathbf{K} \in \mathbb{R}^{p \times p \times m}$ and sample labels $\mathbf{Y} \in \mathbb{R}^{p \times 1}$.
2: **Output**: the learned new kernel $\mathbf{K}_{new}$.
3: Calculate the intra-coupling of base kernels **Intra** using Equation (6).
4: Calculate the inter-coupling of base kernels **Inter** using Equation (9).
5: Learn the base kernel combing coefficient $\boldsymbol{\mu}$ using Equation (21) for binary classification case or Equation (31) for multiple classification case.
6: Generate new kernel $\mathbf{K}_{new}$ using Equation (32) according to the combing coefficient $\boldsymbol{\mu}$.

5 PERFORMANCE EVALUATION

In this section, we evaluate the performance of CMKL-C by comparing it with unweighted sum of kernel methods (UW) and some other state-of-the-art multiple kernel learning methods, including SimpleMKL [5], DBMK-ELM [13], l_1-MK-ELM [22] and radius-incorporated MK-ELM [22]. Considering the computational efficiency and classification performance, we adopt extreme learning machine (ELM) as the classifier in our experiments. Specifically, the output function of ELM with CMKL-C can be written as follows:

$$f(\mathbf{x}) = \begin{bmatrix} k_{new}(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ k_{new}(\mathbf{x}, \mathbf{x}_N) \end{bmatrix}^{\top} \left(\frac{\mathbf{I}}{C} + \mathbf{K}_{new} \right)^{-1} \mathbf{Y} \tag{34}$$

where $\mathbf{Y} = [y_1, \ldots, y_p]^{\top}$ is the label vector, $\mathbf{K}_{new}$ and $k_{new}(\cdot, \cdot)$ is the new training kernel matrix and the new kernel learned by CMKL-C, respectively.

5.1 Benchmark Data Sets

We chose eight classification benchmark data sets including bupa, ionosphere, sonar, wpbc, wine, glass, breast and vowel from UCI Machine Learning Repository [23]. Table 1 shows the number of training samples, testing samples, features and classes

in these data sets. We also evaluate CMKL-C over three bioinformatics benchmark data sets following the experimental setup that was stated in [13]. The first two data sets are about bacterial protein locations [24] while the third data set is the original plant data set of TargetP [25]. The detailed information of these data sets are summarized in Table 2. For each data set, we randomly selected two-thirds of data samples as training data and the remaining part of the data set as testing data. Moreover, we conducted 20 independent trails of experiments for each data set and then compared the average results of different learning methods over the data set.

Data Sets	# train	# test	# features	# classes
bupa	230	115	6	2
ionosphere	234	117	34	2
sonar	138	70	60	2
wpbc	129	65	33	2
wine	118	60	13	3
glass	142	72	9	6
breast	70	36	9	6
vowel	660	330	13	11

Table 1. Summary of the UCI classification data sets

Data Sets	# train	# test	# kernels	# classes
PsortPos	361	180	69	4
PsortNeg	963	481	69	5
plant	627	313	69	4

Table 2. Summary of the bioinformatics benchmark data sets

5.2 Parameters Setting and Evaluation Criteria

Regarding UCI benchmark data sets, 20 Gaussian kernels ($e^{-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2}$) with $\gamma = \{2^{-10}, 2^{-9}, \ldots, 2^9\}$ and 3 polynomial kernels of degree 1, 2 and 3 are used to generate 23 base kernels on full feature vector. Regarding bioinformatics benchmark data sets, we use the same 69 kernels in [12]. Specifically, we use 2 kernels on phylogenetic trees, 3 kernels from BLAST E-values and 64 sequence motif kernels. For all algorithm, the regulation parameter is selected from $\{10^{-1}, 10^0, \ldots, 10^7\}$ via 3-fold cross validation on training data.

In this paper, the classification accuracy is selected as the performance evaluation criteria. We report the results of experiments on each benchmark by using the mean value and the standard deviation of criteria in 20 partitions. We also use the *paired student's t-test* to measure the statistical significance for the accuracy improvement. In *paired student's t-test*, p-value means the probability that

two compared sets coming from the same distribution with equal mean. Typically, if the p-value is less than 0.05, it can be said that the compared sets are having statistically a significant difference.

5.3 Performance Comparison

The classification accuracy of different methods in UCI data sets and bioinformatics data sets are shown in Table 3 and Table 4. The first part of the content in the following table is the *mean* $\pm$ *standard deviation* and the second part is the p-value calculated by the *paired student's t-test*. We denote the highest accuracy and the results that have no significant difference compared to the highest one as bold style.

From the results, CKML-C achieves the highest classification accuracy compared with other state-of-the-art methods. Regarding to UCI benchmarks, CMKL-C significantly outperforms SimpleMKL [5] in the following data sets: *bupa, wpbc, glass, breast* and *vowel*. It is also statistically different from DBMK-ELM [13] in data set *vowel* and has big difference in data sets *glass* and *breast*. Compared with ℓ-MK-ELM and R-MK-ELM [22], CMKL-C appears as significant improvement in most of data set. The result demonstrates the superiority of CKML-C, especially in data sets with multiple classes. It is because CMKL-C considers intra-kernel information and jointly leverages inter-kernel couplings. The intra-kernel information captures the standard deviation in the between-class center direction. It forces the optimal kernel to fit the characteristics of multiple class distribution, thus guarantees the superiority performance when learning multiple class data. The inter-kernel couplings reduce the effect of redundant kernels and enhance the impact of unique information. Therefore, it induces a promising performance for CKML-C compared with other methods. Regarding to the bioinformatics benchmarks, CKML-C dramatically improves state-of-the-art classification performance. On one hand, it shows that the CMKL-C has an appropriate design that drives a better classification performance. On the other hand, it reflects that the bioinformatics benchmarks have strong intra- and inter-kernel couplings, which can feed our proposed CKML-C and should be considered in other analytics tasks.

Overall, the empirical results demonstrate that the proposed CKML-C method overcomes the weakness of existing multiple kernel learning methods, resulting in better classification performance.

6 CONCLUSIONS

In this paper, we have proposed a coupled multiple kernel learning method for supervised classification in the perspective of non-IIDness. The proposed CMKL-C method learns a new kernel by jointly considering intra-coupling within each kernel, inter-coupling among different kernels, and coupling between target labels and real ones. The notable merit of the proposed learning method is that it can fully exploit and fuse information embedded in multiple kernels. Therefore, it can achieve

Data	CMKL-C	SimpleMKL [5]	DBMK-ELM [13]	ℓ_1-MK-ELM [22]	R-MK-ELM [22]	UW
bupa	**68.91 ± 3.68** (**1.00**)	63.78 ± 3.23 (0.00)	**67.48 ± 3.58** (**0.15**)	64.39 ± 3.71 (0.00)	**67.65 ± 3.36** (**0.14**)	**68.13 ± 2.92** (**0.37**)
ionosphere	**95.34 ± 1.37** (**1.00**)	**95.21 ± 1.83** (**0.71**)	**94.96 ± 1.68** (**0.16**)	**95.04 ± 1.77** (**0.38**)	**95.09 ± 1.66** (**0.19**)	94.36 ± 1.70 (0.00)
sonar	**86.21 ± 4.23** (**1.00**)	**85.29 ± 4.89** (**0.25**)	**85.93 ± 3.60** (**0.56**)	85.00 ± 4.00 (0.00)	84.36 ± 4.36 (0.00)	83.00 ± 4.72 (0.00)
wpbc	**79.54 ± 5.02** (**1.00**)	77.15 ± 4.27 (0.00)	**79.31 ± 5.15** (**0.64**)	77.15 ± 4.27 (0.00)	77.31 ± 4.84 (0.00)	78.08 ± 5.18 (0.02)
wine	**98.67 ± 1.59** (**1.00**)	**98.58 ± 1.46** (**0.75**)	**98.42 ± 1.98** (**0.27**)	98.25 ± 1.91 (0.02)	98.25 ± 1.75 (0.02)	97.58 ± 2.51 (0.01)
glass	**68.75 ± 4.74** (**1.00**)	55.97 ± 5.54 (0.00)	**67.99 ± 5.60** (**0.24**)	66.32 ± 4.66 (0.00)	**67.92 ± 4.62** (**0.34**)	66.39 ± 6.46 (0.00)
breast	**70.69 ± 6.71** (**1.00**)	67.08 ± 8.17 (0.01)	**69.31 ± 6.77** (**0.09**)	68.61 ± 6.75 (0.04)	**69.17 ± 6.80** (**0.10**)	**70.14 ± 6.10** (**0.43**)
vowel	**97.70 ± 1.02** (**1.00**)	89.09 ± 1.38 (0.00)	97.14 ± 1.17 (0.00)	96.83 ± 1.07 (0.00)	96.85 ± 1.37 (0.00)	91.20 ± 2.25 0.00
AVG	**73.98**	70.23	73.39	72.40	72.96	72.09

Table 3. UCI benchmarks: Classification accuracy (%). Boldface means no statistical difference from the best one (p-val ≥ 0.05).

Data	CMKL-C	SimpleMKL [5]	DBMK-ELM [13]	ℓ_1-MK-ELM [22]	R-MK-ELM [22]	UW
PsortPos	**88.31 ± 1.99** (**1.00**)	80.22 ± 2.91 (0.00)	87.92 ± 2.03 (0.04)	70.31 ± 3.35 (0.00)	84.14 ± 2.12 (0.00)	81.03 ± 2.69 (0.00)
PsortNeg	**91.81 ± 1.29** (**1.00**)	84.80 ± 1.74 (0.00)	**91.52 ± 0.86** (**0.05**)	73.78 ± 1.82 (0.00)	89.75 ± 1.23 (0.00)	87.31 ± 1.42 (0.00)
plant	**91.88 ± 1.22** (**1.00**)	67.38 ± 3.43 (0.00)	**91.82 ± 1.43** (**0.76**)	58.85 ± 2.96 (0.00)	85.32 ± 2.56 (0.00)	74.79 ± 2.55 (0.00)
AVG	**90.67**	77.47	90.42	67.64	86.40	81.04

Table 4. Bioinformatics benchmarks: Classification accuracy (%). Boldface means no statistical difference from the best one (p-val ≥ 0.05).

higher classification accuracy compared to the existing MKL methods. Numerical results demonstrate that CMKL-C significantly outperforms the state-of-the-art MKL methods in terms of the classification accuracy. In future, we plan to do more in-depth studies regarding non-IIDness based semi-supervised and unsupervised classification/regression methods.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Project No. 61170287, 61232016) and partly by the Natural Science Foundation of Hunan Province (Project No. 2jj3069).

REFERENCES

[1] Harchaoui, Z.—Bach, F.: Image Classification with Segmentation Graph Kernels. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8, doi: 10.1109/cvpr.2007.383049.

[2] Liu, X.—Zhou, L.—Wang, L.—Zhang, J.—Yin, J.—Shen, D.: An Efficient Radius-Incorporated MKL Algorithm for Alzheimer's Disease Prediction. Pattern Recognition, Vol. 48, 2015, No. 7, pp. 2141–2150, doi: 10.1016/j.patcog.2014.12.007.

[3] Vahdat, A.—Cannons, K.—Mori, G.—Oh, S.—Kim, I.: Compositional Models for Video Event Detection: A Multiple Kernel Learning Latent Variable Approach. Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1185–1192, doi: 10.1109/iccv.2013.463.

[4] Lanckriet, G. R.—Cristianini, N.—Bartlett, P.—Ghaoui, L. E.—Jordan, M. I.: Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research, Vol. 5, 2004, pp. 27–72.

[5] Rakotomamonjy, A.—Bach, F.—Canu, S.—Grandvalet, Y.: SimpleMKL. Journal of Machine Learning Research, Vol. 9, 2008, pp. 2491–2521.

[6] Kloft, M.—Brefeld, U.—Sonnenburg, S.—Zien, A.: Lp-Norm Multiple Kernel Learning. Journal of Machine Learning Research, Vol. 12, 2011, pp. 953–997.

[7] Liu, X.—Wang, L.—Yin, J.—Liu, L.: Incorporation of Radius-Info Can Be Simple with Simple MKL. Neurocomputing, Vol. 89, 2012, pp. 30–38, doi: 10.1016/j.neucom.2012.01.035.

[8] Liu, X.—Wang, L.—Yin, J.—Zhu, E.—Zhang, J.: An Efficient Approach to Integrating Radius Information into Multiple Kernel Learning. IEEE Transactions on Cybernetics, Vol. 43, 2013, No. 2, pp. 557–569.

[9] Liu, X.—Wang, L.—Zhang, J.—Yin, J.: Sample-Adaptive Multiple Kernel Learning. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[10] Liu, X.—Yin, J.—Wang, L.—Liu, L.—Liu, J.—Hou, C.—Zhang, J.: An Adaptive Approach to Learning Optimal Neighborhood Kernels. IEEE Transactions on Cybernetics, Vol. 43, 2013, No. 1, pp. 371–384.

[11] Cristianini, N.—Shawe-Taylor, J.—Elisseeff, A.—Kandola, J. S.: On Kernel Target Alignment. Advances in Neural Information Processing Systems, Vol. 14, 2002, pp. 367–373.

[12] Kumar, A.—Niculescu-Mizil, A.—Kavukcuoglu, K.—Daume, H.: A Binary Classification Framework for Two-Stage Multiple Kernel Learning. Proceedings of the 29th International Conference on Machine Learning (ICML 2012), 2012, pp. 1295–1302.

[13] Zhu, C.—Liu, X.—Liu, Q.—Ming, Y.—Yin, J.: Distance Based Multiple Kernel ELM: A Fast Multiple Kernel Learning Approach. Mathematical Problems in Engineering, Vol. 2015, 2015, Art. No. 372748, 9 pp., doi: 10.1155/2015/372748, doi: 10.1155/2015/372748.

[14] Cao, L.: Coupling Learning of Complex Interactions. Information Processing and Management, Vol. 51, 2015, No. 2, pp. 167–186.

[15] Cao, L.: Non-IIDness Learning in Behavioral and Social Data. The Computer Journal, Vol. 57, 2014, No. 9, pp. 1358–1370.

[16] Cao, L.—Ou, Y.—Yu, P. S.: Coupled Behavior Analysis with Applications. IEEE Transactions on Knowledge and Data Engineering, Vol. 24, 2012, No. 8, pp. 1378–1392.

[17] Wang, C.—Cao, L.—Wang, M.—Li, J.—Wei, W.—Ou, Y.: Coupled Nominal Similarity in Unsupervised Learning. Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 973–978, doi: 10.1145/2063576.2063715.

[18] Wang, C.—She, Z.—Cao, L.: Coupled Attribute Analysis on Numerical Data. Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 1736–1742.

[19] Wang, C.—She, Z.—Cao, L.: Coupled Clustering Ensemble: Incorporating Coupling Relationships Both Between Base Clusterings and Objects. Proceedings of the IEEE International Conference on Data Engineering, 2013, pp. 374–385, doi: 10.1109/icde.2013.6544840.

[20] Nguyen, C. H.—Ho, T. B.: An Efficient Kernel Matrix Evaluation Measure. Pattern Recognition, Vol. 41, 2008, No. 11, pp. 3366–3372, doi: 10.1016/j.patcog.2008.04.005.

[21] Boyd, S.—Vandenberghe, L.: Convex Optimization. Cambridge University Press, 2004.

[22] Liu, X.—Wang, L.—Huang, G.-B.—Zhang, J.—Yin, J.: Multiple Kernel Extreme Learning Machine. Neurocomputing, Vol. 149, 2015, pp. 253–264, doi: 10.1016/j.neucom.2013.09.072.

[23] Bache, K.—Lichman, M.: UCI Machine Learning Repository. Available on: `http://archive.ics.uci.edu/ml`, 2013.

[24] Gardy, J. L.—Laird, M. R.—Chen, F.—Rey, S.—Walsh, C.—Ester, M.—Brinkman, F. S.: PSORTb v.2.0: Expanded Prediction of Bacterial Protein Subcellular Localization and Insights Gained from Comparative Proteome Analysis. Bioinformatics, Vol. 21, 2005, No. 5, pp. 617–623, doi: 10.1093/bioinformatics/bti057.

[25] Emanuelsson, O.—Nielsen, H.—Brunak, S.—von Heijne, G.: Predicting Subcellular Localization of Proteins Based on Their n-Terminal Amino Acid Sequence. Journal of Molecular Biology, Vol. 300, 2000, No. 4, pp. 1005–1016, doi: 10.1006/jmbi.2000.3903.

En Zhu received his M.Sc. and Ph.D. degrees in computer science from National University of Defense Technology, Changsha, China, in 2001 and 2005, respectively. He was awarded the National Excellent Doctoral Dissertation in 2007. He is Professor with the School of Computer Science, National University of Defense Technology, Changsha, China. From 2009 to 2010, he visited the Department of Computer Science, University of York, York, U.K. His research interests include pattern recognition, image processing, machine vision, and machine learning.

Qiang Liu received his Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT) in 2014. He has contributed several archived journal and international conference papers, such as IEEE Transactions on Wireless Communications, IEEE Communications Letters, Neurocomputing, Neural Computing and Applications, etc. He was invited as a TPC member of several international conferences, e.g. ITNAC '15, SUMS '15, CHINACOM '14, QShine '15 – SNCC Workshop, and as a Session Chair of HPCC '13. He is a member of IEEE/ACM and a member of China Computer Federation (CCF). His research interests include protocol design and performance evaluation, machine learning, denial-of-service detection as well as other security issues in emerging wireless networks.

Jianping Yin received his Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT), Changsha, China, in 1990. He is Professor of the College of Computer, NUDT and currently holds the positions of the Dean of the Department of Computer Science and Technology. In the past, he was the head of China Computer Federation Technical Committee on Theoretical Computer Science. His research interests include artificial intelligence, network algorithm and information security.