Ensemble Kernel Ridge Regression from a Multi-view Perspective

ZHIFENG LIU, School of Computer Science and Communication Engineering, Jiangsu University, China
LIU CHEN, CONGHUA ZHOU, SUMET MEHTA, and XIANG-JUN SHEN*, School of Computer Science and Communication Engineering, Jiangsu University, China
YU-BAO CUI, Department of Clinical Research Center, The Affiliated Wuxi People's Hospital of Nanjing Medical University, China

In this paper, we develop an ensemble kernel ridge regression (E-KRR) model from the traditional linear ridge regression in multi-view assumption to improve the performance of regression and classification tasks. Motivated by multi-view data modeling, in our proposed ensemble framework, original data samples are assumed virtually to have multi-view presentations. Then these virtual multi-view presentations are modeled in the traditional ridge regression method. With the appropriate deduction, we then transform our multi-view linear ridge regression model into a multiple ensemble kernel ridge regression model, where virtual multi-view presentations become multi-kernel representations. Also, their corresponding regression parameters in linear space are transformed into the corresponding kernel parameters in multiple kernel spaces. Therefore, the problem of the kernel and its parameter selection in traditional single KRR method is overcome by finding the best combinational kernel representations and their parameters in multiple Reproducing Kernel Hilbert Spaces (RKHSs), which results in a better overall regression performance. Experimental results on various regression and classification datasets demonstrate that the proposed method significantly outperforms the other state-of-the-art regression and ensemble methods, such as XGBoost, GBDT, and GBDT-PL.

CCS Concepts: • **Computing methodologies** → **Kernel methods**; **Ensemble methods**; *Classification and regression trees.*

Additional Key Words and Phrases: ensemble learning, kernel ridge regression, multiple kernel learning, multi-view

ACM Reference Format:
Zhifeng Liu, Liu Chen, Conghua Zhou, Sumet Mehta, Xiang-Jun Shen, and Yu-bao Cui. XX. Ensemble Kernel Ridge Regression from a Multi-view Perspective. *ACM Trans. Intell. Syst. Technol.* xx, xx, Article xxx (July XX), 19 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Ridge regression (RR) is the most elementary algorithm that can be kernelized and is a widely accepted efficient machine learning paradigm, which has been well used in solving regression [11],

*Corresponding author. E-mail address: xjshen@ujs.edu.cn

Authors' addresses: Zhifeng Liu, School of Computer Science and Communication Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang, Jiangsu, China, 212013, liuzf@ujs.edu.cn; Liu Chen, 278990214@qq.com; Conghua Zhou, chzhou@ujs.edu.cn; Sumet Mehta, msumet@outlook.com; Xiang-Jun Shen, School of Computer Science and Communication Engineering, Jiangsu University, No. 299 at Qingyang Road, Wuxi, Jiangsu, China, 214023; Yu-bao Cui, ybbcui1975@hotmail.com, Department of Clinical Research Center, The Affiliated Wuxi People's Hospital of Nanjing Medical University, 301 Xuefu Road, Zhenjiang, Jiangsu, China, 212013.

2157-6904/XX/7-ARTxxx $15.00
https://doi.org/XXXXXXX.XXXXXXX

[20] and classification [25] [22] [13] problems. Such as Sun et al. [31] proposed a novel clustering algorithm based on local learning, which uses kernel regressor as a local label predictor. He et al. [15] proposed a kernel ridge regression classification algorithm (KRC) based on ridge regression. KRC algorithm first maps the data nonlinearly to the feature space, and then carries out ridge regression classification on the feature space. But, this method requires matrix inversion calculation, resulting in a large amount of calculation. Also, Weinberger and Tesauro [35] proposed metric learning for kernel regression (MLKR) method to find the optimal linear subspace with the minimum Nadaraya-Watson square error in the training set of a common Gaussian kernel.

Some researchers have combined kernel learning with ensemble learning due to the good performance of ensemble learning. The basic principle of the ensemble method is to integrate multiple basic models to build regressor. The ensemble regressor is better than the single regression model prediction. For instance Tsai et al. [32] proposed a clustering-based feature selection method, which clusters images with multiple feature representations, integrates multi-kernel learning into the training process of self-organizing mapping, and associates each cluster with a learnable ensemble kernel. Through optimization iteration, the quality of the ensemble kernel is gradually improved by strengthening the clustering structure.

However, the performance of kernel-based regression models mainly depends on the choice of the kernel function. Using a single kernel function to train regression or classification model is prone to over-fitting or under-fitting, resulting in low generalization ability and poor robustness. At the same time, the use of an inappropriate kernel function may lead to the degradation of model performance, so it is difficult to determine the optimal combination kernel function in practical application.

To solve the problem of kernel and its parameter selection of a kernel model, researchers have done a lot of research. Such as Samah et al. [24] proposed a technique to eliminate false edges from binary edge images by using a local adaptive regression kernel as the descriptor of edge detection. Livetsky [18] extended the single parameter ridge regression model to the two-parameter model and obtained a variety of asymptotic behaviors with better fitting characteristics. Salhov et al. [23] designed the kernel by approximating the similarity between the ensemble parameters shared by multiple feature subsets. In [42], Zhang et al. showed that multi-task extension of Multi-Kernel Support Vector Machines improves the detection of Action Units. In [27] [38] [4], multiple kernels with different types of features are used instead of a single kernel with a single type of features, which improves the performance of the support vector machine (SVM) in facial expression and human motion recognition.

Some researchers have also applied multi-view learning to optimize the above problems. Benedict et al. [8] proposed a web server based on unsupervised multi-kernel learning for multi-view data dimensionality reduction and sample clustering. Yuan et al. [1] proposed a framework for multi-view image recognition. Its core idea is to map multiple views to multiple high-dimensional feature spaces through multiple nonlinear mappings determined by different kernels. This method can find a variety of useful information of each original view in the feature space. Ye et al. [40] found potential clusters by maximizing the weighted similarity between individual view clusters based on a k-means algorithm. ./output.bbl

Based on the above related research on KRR and inspired by multi-view data processing, we propose a combined kernel function and develop an ensemble kernel ridge regression (E-KRR) model based on the traditional linear ridge regression under the assumption of multi-view data. In our proposed ensemble framework, as shown in Figure 1, it is assumed that the original data samples have multi-view representation. Through multi-view data representation, we develop an ensemble kernel ridge regression method directly from the linear ridge model. Then, the multi-view representation in the original data is transformed into multiple kernel representations. The

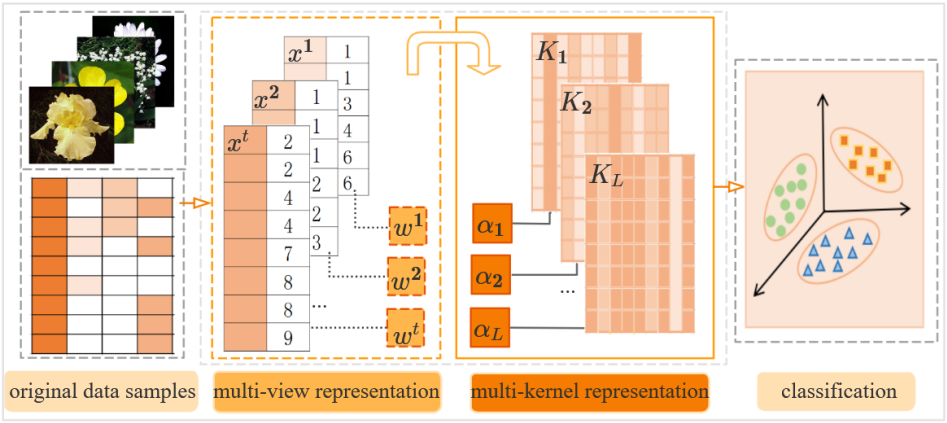


Fig. 1. Motivational diagram of our ensemble kernel ridge regression.

corresponding regressor or classifier parameters in linear space are transformed into corresponding kernel parameters in multi-kernel space. Meanwhile, each kernel regressor or classifier factor in multi-kernel space is associated with a weight, which comes directly from the data view without any manual intervention. Thus, the problem of kernel and parameter selection in the traditional single KRR method is overcome. In this way, our ensemble kernel ridge regression and classification model has the advantages of both global kernel function and local kernel function.

The main contributions of this paper are as follows:

1) First, motivated by multi-view data modeling idea, we develop a novel ensemble kernel ridge regression model. In our proposed ensemble model, original data samples are assumed to have multi-view presentations and therefore multi-view presentations in original data can be transformed into multiple kernel representations. That means multi-view data assumptions now can be represented by multiple kernel representations, and their corresponding regressor or classifier parameters in linear space are transformed into corresponding kernel parameters in multiple kernel spaces.
2) Second, our proposed ensemble model can overcome the problem of the kernel and its parameter selection in the traditional single KRR method by finding the best combinational kernel representations and their parameters in multiple Reproducing Kernel Hilbert Spaces (RKHSs). In this way, the best combinational kernel representations and their parameters can be found in multiple Reproducing Kernel Hilbert Spaces (RKHSs) and therefore a new ensemble regression and classification model with good performance is obtained.
3) Finally, we verify that our proposed ensemble model is superior to other most advanced regression and ensemble methods in classification and regression tasks.

The structure of this paper is as follows: Section 2 introduces the related work of kernel ridge regression and multi-view date processing. In Section 3, the proposed Ensemble Kernel Ridge Regression from a Multi-view Perspective is introduced. In Section 4, experiments are carried out on UCI and different image datasets respectively. Finally, we summarize this paper in the section 5.

2 RELATED WORK

In regression problem, it is difficult to establish a regression model with excellent and stable performance in all aspects. Usually, we can only get a regression model with multiple preferences. The emergence of ensemble learning overcomes this difficulty, and the ensemble learning model

is more accurate than the single model. Ensemble learning (EL) completes the learning task by constructing and merging multiple weak regression classifiers, that is, to improve the accuracy and diversity of base regression classifiers as much as possible. Common ensemble learning algorithms include XBGoost, random forest (RF), etc. The idea of ensemble learning appeared at the end of the 20th century. Until 2003 mukkamala et al. [19] applied the idea of ensemble learning to the field of intrusion detection and fused two different classifiers, artificial neural network and support vector machine (SVM). In the next period of time, a large number of ensemble learning methods were proposed [26] [36]. Such as in literature [26], they preprocessed the data by using the ensemble learning method, and then combined the models learned by Bayesian network and classification regression tree.

At present, most studies use logistic regression and SVM as basis regression classifiers for ensemble learning [6] [21]. Such as in [2], ten SVM classifiers are used as members of the credit scoring ensemble model. In [39], Yao et al. proposed a new hybrid RF-SVM ensemble model, which uses random forest to select important variables, and uses ensemble methods (bagging and boosting) to aggregate a single base model (SVM) as a robust classifier. However, linear classifiers such as logistic regression and SVMs are stable classifiers, which are not sensitive to sample disturbance. Therefore it is very difficult to build a variety of base classifiers through sample disturbances.

Moreover, many researchers use multiple regression classifiers to form heterogeneous ensemble learning models [34], [3]. Such as in [34], considering the large unbalanced data, Wang et al. used regular logistic regression as the basic classifier. Firstly, the data is balanced and diversified by clustering and packaging algorithms. Then lasso logistic regression learning ensemble is applied to evaluate credit risk. However, the dimensions of prediction probability output by different types of classifiers are different, which can-not be averaged directly, resulting in poor regression classification results.

At the same time, the tree model is an unstable regression classifier with sample disturbance. It is easy to construct a variety of regression classifiers through sample disturbance. Therefore, the tree model is a very popular basic model in ensemble methods. Gradient boosting decision trees (GBDT) is a kind of tree model and a common nonlinear model. The goal of each iteration is to reduce the gradient direction of the residual. This idea enables GBDT to find various features for feature combinations. Such as Li et al. [16] established a gradient boosting decision tree model to predict the mortality of COVID-19. Feng et al. [**?**] developed a learning architecture LR2GBDT for stock index prediction and trading by cascading the logistic regression (LR) model to the GBDT model. In [28], Shi et al. further improved the accuracy and efficiency of GBDT by using more complex basic learners. The extended gradient lifting uses a piecewise linear regression tree (PL tree) as the basic learner instead of piecewise constant regression tree. In addition XGBoost is a gradient lifting algorithm based on ensemble learning, and it is also a kind of tree model. Based on logistic regression and XGBoost, diabetes prediction models [36] were established respectively. Experiments of diabetes prediction show that the prediction accuracy of XGBoost is higher than that of traditional logistic regression.

However, using a decision tree as the basic learner makes a large prediction error when separating boundary points. Some studies use kernel ridge regression instead of the tree model. Such as Sun et al. [30], used KRR to replace the decision tree in random forest and boosting method. The simulation results of prostate cancer and Boston housing data show the effectiveness of the ensemble method with kernel ridge regression. Also, Zhang and Suganthan [41] extended the work on oblique decision tree ensemble and proposed an effective joint training kernel ridge regression method. Belle and Lisboa [7] proposed a new flexible sparse classifier for interpretable decision support system, and extended RBF kernel to interpretable and visual components.

In general, during dividing the original problem, some edge information or overall information will be lost. Most of the ensemble learning algorithms doesn't consider the aforementioned situation, which will have impact on the final results. Moreover, setting different parameters for the regressor and classifier will directly affect its final effect, so the process needs a lot of debugging. Form the above discussions, we can see that there is still some room for improvement in the ensemble learning algorithms for regression and classification.

3 THE PROPOSED APPROACH

Multi-view data processing is to describe the same data from different perspectives. Inspired by the traditional linear ridge regression and multi-view data processing, we propose an ensemble kernel ridge regression model. It is assumed that the original data sample has a multi-view representation. The multi-view representation of the original data in the linear kernel ridge regression model is transformed into multiple kernel representations in the ensemble kernel ridge regression model. At the same time, the regression parameters in linear space are transformed into the corresponding kernel parameters in multi-kernel space. Therefore, each kernel regressor factor in multi- kernel space is directly associated with the weight in the data view, which solves the problem of kernel and its parameter selection in the traditional single KRR method.

3.1 Preliminary

In some practical problems, there are many uncorrelated or redundant features among high-dimensional data samples, which bring great challenges to the traditional learning algorithms. At the same time, in many practical problems, there are multiple views of data, and the labels of data are difficult to obtain, so multi-view learning was proposed to overcome these difficulties. Suppose x_i represents the i^{th} data sample and $x^{(t)}$ represents the t^{th} view of the data, the multi-view data can be represented as $x = \left\{x_i^1, x_i^2, ..., x_i^L\right\}$. The research on multi-view data is called multi-view learning, which can find out the important information contained in the data. Therefore, under the assumption of multi-view data, we assume that the original data samples have multi-view representation, and then directly develop an ensemble kernel ridge regression method from the linear ridge model.

Ridge regression is the basic kernel method, whereas KRR extends the ridge regression into nonlinear cases and uses kernel trick to solve nonlinear separable problems. A typical linear regression problem can be expressed as:

$$\min_w \sum_i \left(w^T x_i - y_i\right)^2 + \lambda \|w\|^2 \tag{1}$$

where the parameter λ is a user-defined regularization parameter, which is used to control the complexity of the model. This problem can be transformed into a closed-form solution:

$$w = \left(X^T X + \lambda I_D\right)^{-1} X^T y \tag{2}$$

where each row of the data matrix X has a sample. I is an identity matrix. Each element of the vector y_i is the output target of x_i.

Further, based on the matrix inverse lemma [5], the above inversion can be performed in the smallest space of the two possibilities (the dimension of the feature space or the number of data cases) and finally the solution of w can be expressed as a linear combination of samples in the feature space $\phi(x)$ as $w = \sum_i \phi(x_i)\alpha_i$ with $\alpha = (K + \lambda I_N)^{-1} Y$.

Therefore, KRR extends linear regression nonlinearly through kernel trick. By using the kernel trick, the kernel matrix K can be obtained by $K_{i,j} = k\left(x_i, x_j\right) = \phi(x_i)^T \phi\left(x_j\right)$. And thus we can get

$y = w^T \phi(x_i) = \sum_i \alpha_i K(x_i, x_j)$. Finally, the KRR problem is converted to the following formula:

$$\min_{\alpha} \|K\alpha - Y\|^2 + \lambda \alpha^T K \alpha \tag{3}$$

3.2 Proposed ensemble kernel ridge regression model

Based on the above discussions, it is assumed that the original data samples have multi-view representation. The multi-view representation in the original data is transformed into multiple kernel representations in multi-kernel space, and the regression parameters in linear space are transformed into the corresponding kernel parameters in multi-kernel space. Thus, the optimal combination of kernel representation and its parameters are found in multiple Reproducing Kernel Hilbert Spaces (RKHSs).

Suppose in a regression problem, training set $X = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, and testing set $X_t = \{(x_1, x_2, ..., x_{N_t})\}$. Also x_n $(x_n \in R^d, n = 1, ..., N)$ represents a training sample, x_t $(x_t \in R^d, t = 1, ..., N_t)$ represents a testing sample, y_n is the regression result of x_n, and N is the number of training samples while N_t is the number of testing samples. According to formula (1), a single kernel ridge regression model is as follows:

$$f(H) = \|H - Y\|_F^2 \tag{4}$$

where $H = X^T w + b \cdot 1$ is a variable of a real function $f(H)$, X is sample feature. w is the weight of the regressor. b is the deviation of the regressor, and Y is the output value.

Formula (4) is the traditional ridge regression model and it is a convex optimization problem. However, the following Jensen's inequality [14] give us a modeling idea on building a more compact regression model. For a real function $\phi(x)$ when $\phi(x)$ is a convex function in interval I, it satisfies the following relationship,

$$\begin{gathered} \phi\left(\sum_{i=1}^L p_i x_i\right) \leq \sum_{i=1}^L p_i \phi(x_i) \\ S.T.\ p_i \geq 0, \sum_{i=1}^L p_i = 1, x_i \in I\ (i = 1, ..., L) \end{gathered} \tag{5}$$

which is called the Jensen's Inequality [14]. Inspired by Jensen's Inequality, we take each regressor H_i as a variable of a real function $f(H_i)$, and thus for the real function $f(H_i)$ of each kernel regressor, we can draw an analogy from formula (5),

$$\begin{gathered} f\left(\sum_{i=1}^L p_i H_i\right) = \left\|\sum_{i=1}^L p_i H_i - Y\right\|_F^2 \leq \sum_{i=1}^L p_i f(H_i) \\ S.T.\ p_i > 0, \sum_{i=1}^L p_i = 1, H_i \in I\ (i = 1, ..., L) \end{gathered} \tag{6}$$

Therefore, through formula (6) and Jensen's Inequality, it can conduct our compact E-KRR model since the right side of the above formula has a global optimal solution, where it is a weighted combinational convex regression. Meanwhile, considering the problem of overfitting, a regular term $\sum_{i=1}^L w_i^T w_i$ are added. Finally, our proposed E-KRR model is as follows:

$$\begin{gathered} \underset{p, w_i}{\arg\min} \frac{1}{2} \left\|\sum_{i=1}^L p_i H_i - Y\right\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^L w_i^T w_i \\ S.T.\ H_i = X_i^T w_i + b_i \cdot 1, 1^T p = 1, p > 0 \end{gathered} \tag{7}$$

where H_i is the representation of the data in the i^{th} view and $H = [H_1 H_2 ... H_L]$, L is the number of base regressors. p_i is the regression coefficient of the i^{th} regressor, with the size of $L \times 1$. X_i is the i^{th} sample feature. w_i is the weight of the i^{th} regressor, with the size of $N \times 1$. b_i is the deviation of the i^{th} regressor. Y is the output of the i^{th} sample, and λ is the penalty parameter.

3.3 Optimization

In this section, the solution process of our proposed E-KRR model is described in detail. We convert the multi-view representation in the original data into multiple kernel representations. Then, the corresponding regression parameters in linear space are transformed into the corresponding kernel parameters in multi-kernel space. Since our E-KRR model is an optimization problem, we introduce the augmented Lagrange multiplier (ALM) [17] method to solve the optimal solution of formula (7). ALM method is used to solve optimization problems under equality constraints. By introducing Lagrange multiplier $\lambda, \rho_i, \eta, \xi, \mu$, formula (7) can be transformed into the following unconstrained problem (8):

$$\begin{aligned} &\underset{w_i,p,H_i,b_i}{\arg\min} L\left(w_i, p, H_i, b_i, \lambda, \rho_i, \eta, \xi, \mu\right) \\ &= \frac{1}{2}\left\|\sum_{i=1}^{L} p_i H_i - Y\right\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{L} w_i^T w_i + \sum_{i=1}^{L}\left\langle \rho_i, X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right\rangle \\ &+ \left\langle \eta, 1 - 1^T p\right\rangle + \left\langle \xi, p\right\rangle + \frac{\mu}{2}\left(\sum_{i=1}^{L}\left\|X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right\|_2^2 + \left\|1 - 1^T p\right\|_2^2 + \|p\|_2^2\right) \end{aligned} \tag{8}$$

Solution to parameter w_i

First, find the closed form solution of w_i. The solution problem of variable w_i is transformed into its effective sub-problem, and equation (8) is rewritten as:

$$\begin{aligned} J\left(w_i\right) &= \frac{\lambda}{2}\sum_{i=1}^{L} w_i^T w_i + \sum_{i=1}^{L}\left\langle \rho_i, X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right\rangle + \frac{\mu}{2}\left(\sum_{i=1}^{L}\left\|X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right\|_2^2\right) \\ &= \frac{\lambda}{2}\sum_{i=1}^{L} w_i^T w_i + \sum_{i=1}^{L}\rho_i^T X_i^T w_i + \frac{\mu}{2}\sum_{i=1}^{L}\left[2(b_i \cdot 1_{N*1} - H_i)^T X_i^T w_i + w_i^T X_i X_i^T w_i\right] \end{aligned} \tag{9}$$

Second, taking the derivative of w, this problem can be transformed into a closed-form solution:

$$w_i = \left(\frac{\lambda}{\mu} I + X_i X_i^T\right)^{-1} X_i\left(-\frac{1}{\mu}\rho_i - b_i \cdot 1_{N*1} + H_i\right) \tag{10}$$

where X_i is a data representation. Based on the matrix inverse lemma [5], the above inversion can be performed in the smallest space of the two possibilities, and equation (10) is rewritten as:

$$\hat{w}_i = \phi\left(\phi^T\phi + \frac{\lambda}{\mu} I_N\right)^{-1}\left(-\frac{1}{\mu}\rho_i - b_i \cdot 1_{N*1} + H\right) \tag{11}$$

Then the solution of $\hat{w}_i$ can be expressed as a linear combination of samples in the feature space $\phi_i\left(x\right)$ as $\hat{w} = \sum_{j=1}^{N}\phi_i\left(X_j\right)\alpha_i$ as

$$\alpha_i = \left(\phi_i^T\phi_i + \frac{\lambda}{\mu} I_N\right)^{-1}\left(-\frac{1}{\mu}\rho_i - b_i \cdot 1_{N*1} + H_i\right) \tag{12}$$

where ϕ_i is the kernel representation of the i^{th} data. By using the kernel trick $K_i = k\left(x_j, x_k\right) = \phi_i\left(x_j\right)^T\phi_i\left(x_k\right)$, we have our ensemble regression as:

$$Y = \sum_{i=1}^{L} p_i H_i = \sum_{i=1}^{L} p_i\left[K_i\alpha_i + b_i \cdot 1\right] \tag{13}$$

Therefore we convert our virtual multi-view representation into multiple kernel representations. **Therefore, based on the multi-view data representation, we directly develop our E-KRR method from the linear ridge model and convert the multi-view representation in the original data into multiple kernel representations.** Now as each virtual data X_i is converted into kernel representation, the other parameters are updated as follows:

Update parameter H_i

First, find the closed form solution of H_i. The solution problem of variable H_i is transformed into its effective sub-problem, and equation (8) is rewritten as:

$$\begin{aligned} J(H_i) &= \frac{1}{2}\left(\sum_{i=1}^{L} p_i H_i - Y\right)^T \left(\sum_{i=1}^{L} p_i H_i - Y\right) + \sum_{i=1}^{L} \rho_i^T \left(X_i^T w_i + b_i \cdot 1_{N1} - H_i\right) \\ &\quad + \frac{\mu}{2} \sum_{i=1}^{L} \left(X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right)^T \left(X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right) \\ &= \frac{1}{2} \sum_{i=1}^{L} \left(H_i^T p_i p_i H_i - 2 p_i H_i^T Y\right) - \sum_{i=1}^{L} \rho_i^T H_i + \frac{\mu}{2} \left[\sum_{i=1}^{L} H_i^T H_i - 2\left(X_i^T w_i + b_i \cdot 1_{N1}\right)^T H_i\right] \end{aligned} \tag{14}$$

Second, transform the regression parameter H_i in linear space into the corresponding kernel parameter $\hat{H}_i$ in multi-kernel space. Based on formula (12), we can obtain the updating solution to $\hat{H}_i$ by setting the partial derivative of $J(H_i)$ to zero :

$$\hat{H}_i = \frac{1}{(\mu + p_i^2)} \left[Y p_i - p_i \sum_{j=1, j \neq i}^{L} H_j p_j + \rho_i\right] + \frac{1}{(\mu + p_i^2)} \mu \left(K_i \alpha_i + b_i \cdot 1_{N*1}\right) \tag{15}$$

Thus, the regressor parameter H_i represented by multi-view are transformed into the kernel regressor parameter $\hat{H}_i$ represented by multi-kernel representation.

Update parameters p and b_i

The corresponding regression parameters in linear space also need to be transformed into the corresponding kernel parameters in multi-kernel space. We convert the solution problem of variables p and b_i to its sub-problem as:

$$J(p) = \frac{1}{2}\left(p^T H^T H p - 2 p^T H^T Y\right) - \eta 1^T p + \xi^T p + \frac{\mu}{2}\left[-1_{L*1}^T p - p^T 1_{L*1} + p^T 1_{L*1} 1_{L*1}^T p + p^T p\right] \tag{16}$$

$$J(b_i) = \sum_{i=1}^{L} \rho_i^T \left(X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right) + \frac{\mu}{2} \sum_{i=1}^{L} \left(X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right)^T \left(X_i^T w_i + b_i \cdot 1_{N*1} - H_i\right) \tag{17}$$

Then, taking the derivative of p and b_i respectively, we can get the result of parameters p and $\hat{b}_i$ in kernel space :

$$p = \left(H^T H + \mu \cdot 1_{L*L} + I\right)^{-1} \left(H^T Y + \eta \cdot 1_{L*1} - \xi + \mu \cdot 1_{L*1}\right) \tag{18}$$

$$\hat{b}_i = \frac{1}{N}\left[1_{N*1}^T \left(K_i \alpha_i - H_i\right) - \frac{1}{\mu}\left(\rho_i^T 1_{N*1}\right)\right] \tag{19}$$

Finally, in order to find the best combination kernel representation and its parameters in multiple Reproducing Kernel Hilbert Spaces (RKHSs), we update all the above Lagrange multipliers as follows, where $\theta_1, \theta_2, \theta_3, \theta_4$ are the learning rate:

$$\rho_i := \rho_i + \theta_1 \left(K_i \alpha_i + b_i \cdot 1_{N*1} - H_i\right) \tag{20}$$

$$\eta := \eta + \theta_2 \left(1 - 1^T p\right) \tag{21}$$

$$\xi := \xi + \theta_3 p \tag{22}$$

$$\mu := \mu + \frac{\theta_4}{2}\left(\sum_{i=1}^{L} \left\|K_i \alpha_i + b_i \cdot 1_{N*1} - H_i\right\|_2^2 + \left\|1 - 1^T p\right\|_2^2 + \|p\|_2^2\right) \tag{23}$$

Repeat this process iteratively until the loss converges. The details are shown in Algorithm 1.

Algorithm 1 Ensemble Kernel Ridge Regression from a Multi-view Perspective.

1: Input: Training data, kernel function parameter.
2: Initialize: $\xi = zeros\left(L,1\right), \rho_i = zeros\left(N,1\right), \mu = e^{-9}$
3: **while** loss not converged **do**
4: Update w_i, H_i, p, b_i by Eq.(10), Eq.(15), Eq.(18) and Eq.(19).
5: Update ρ_i, η, ξ and μ by Eq.(20)-Eq.(23).
6: Check convergence condition till the loss is converged or the maximum iteration number is reached.
7: **end while**
8: Output: α_i, p_i, b_i

3.4 Analysis on computational complexity

In this section, we concentrate on the computational complexity of the proposed algorithm above. The time complexity of the parameter p is $\mathrm{O}\left(L^3\right)$ due to the L-dimensional matrix needs to be inverse. The time complexity of coefficients α is $\mathrm{O}\left(N^3\right)$ in each iteration, where N is the number of sampls and K is the number of sample features. There are L coefficients that need to be updated, which makes a time complexity of $\mathrm{O}\left(LN^3\right)$. Assuming that the number of iterations are C, so the time complexity of the loop update phase is $\mathrm{O}\left(LN^3\right)$. Since $N \gg C$, $N \gg L$ and $N \gg K$, the overall time complexity of our proposed algorithm is $\mathrm{O}\left(N^3\right)$.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we have conducted experiments for the regression and classification problems on UCI datasets to verify the performance of our proposed ensemble kernel ridge regression method. Firstly, the parameter setting is discussed in Section 4.1, and the results of regression and classification experiments are performed and analyzed in detail in sections 4.2 to 4.3 respectively.

4.1 Parameter Setting

In order to enhance the reliability of the regression and classification experimental results, we randomly divide the dataset into 70% for training and 30% for testing. Then the regularization parameters are obtained by cross validation. We select six comparative methods to show the superiority of our proposed E-KRR method, including Random forest (RF) [37], eXtreme Gradient Boosting (XGBoost) [10], Gradient Boosting Decision Tree (GBDT) [33], GBDT-PL [29], gcF [44] and MLS-SVR [45]. In our proposed method, a single kernel regressor model is used as the basic model of ensemble kernel ridge regression such as polynomial kernel: $k\left(x_i, x_j\right) = \left(ax_i^T x_j + c\right)^d$, radial basis function kernel model: $k\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right)$ and gaussian kernel model: $k\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$. The kernel parameters of the single kernel regressor model are transformed from the corresponding parameters in the multi-view representation. We focus on how to combine basic kernel regressors from the kernel pool. Therefore, we choose Mean Square Error (MSE) and Mean Absolute Error (MAE) as the evaluation criteria,

$$MSE = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(f\left(x_i\right) - y_i\right)^2 \tag{24}$$

$$MAE = \frac{1}{N_t}\sum_{i=1}^{N_t}\left|f\left(x_i\right) - y_i\right| \tag{25}$$

Table 1. Description of UCI datasets.

Datasets	Samples	Features
SB	8760	13
Bi	7752	23
RQ	1599	12
Mg	1385	7
FF	517	9
He	267	45
Cp	8192	12
Sp	3107	6

where N_t is the size of testing data, $f(x_i)$ is the predicted value of the i^{th} sample while y_i is the corresponding true value.

For regression experiments, there are three parameters a, c, d for the polynomial kernel, and one parameter σ for radial basis function kernel, gaussian kernel, exponential kernel, multi-quadric kernel and laplacian kernel. Different settings of the parameters will give different performance on the experimental results. Generally, we set $a \in \{1*1e-6, 1*1e-5, ..., 1000\}$, $c \in \{1*1e-6, 1*1e-5, ..., 1000\}$, $d \in \{1, 2, 3, 4, 5\}$ and $\sigma \in \{1*1e-6, 1*1e-5, ..., 1000\}$. Besides, we make the number of leaf nodes same with the number of kernels in our ensemble methods, such as RF and XGBoost.

For classification experiments, we use six comparative classification methods, which are Random Forest (RF) [37], eXtreme Gradient Boosting (XGBoost) [10], Gradient Boosting Decision Tree (GBDT) [33], KOC+ [9], MENLR [43] and MLS-SVR [45]. In order to use kernel method, we change the kernel in the regression method, and then reset $a \in \{1*1e-4, 1*1e-3, ..., 1000\}$, $c \in \{1*1e-4, 1*1e-3, ..., 1000\}$, $d \in \{1, 2, 3, 4\}$ and $\sigma \in \{1*1e-4, 1*1e-3, ..., 1000\}$ accordingly.

For each dataset, the optimal parameters and basic kernel models are obtained by 10-fold cross validation in experiments. The parameter L in Eq. (7) denotes the number of basic kernel models, and we set $L \in \{20, 40, 70, 100\}$ due to proper selection of L can show good generalization ability.

4.2 Regression on UCI datasets

We select eight public datasets from UCI datasets repository for validation of our proposed E-KRR model, including SeoulBike (SB), Biascorrection (Bi), RedWineQuality (RQ), Mg, ForestFires (FF), Heart (He), Cpusmall (Cp) and Space_ga (Sp). A detail description of these datasets is presented in Table 1.

We discuss the general performance of the proposed ensemble kernel ridge regression algorithm and the six comparative methods, including RF [37], XGBoost [10], GBDT [33], GBDT-PL [29], gcF [44] and MLS-SVR [45].

Table 2 shows the Mean Squared Error (MSE) results of our proposed method and other models on eight UCI datasets. We can see that our regression method performs better than other single models or ensemble models. From the results, all methods performed well on SB, Mg, and Sp datasets, where, the performance on the Sp dataset is the best. However, our method has the best performance, with values of 0.0718, 0.0152 and 0.0083 respectively, which are 1.1% (gcF), 3.7% (GBDT-PL) and 11.4% (g GBDT-PL) lower than the second-best method, respectively. For Bi, FF and CP datasets, the performance of our proposed method is much better than other methods. Compared with the worst performing method, it reduces 3.4128 (RF), 3.8877 (GBDT-PL) and 6.2121

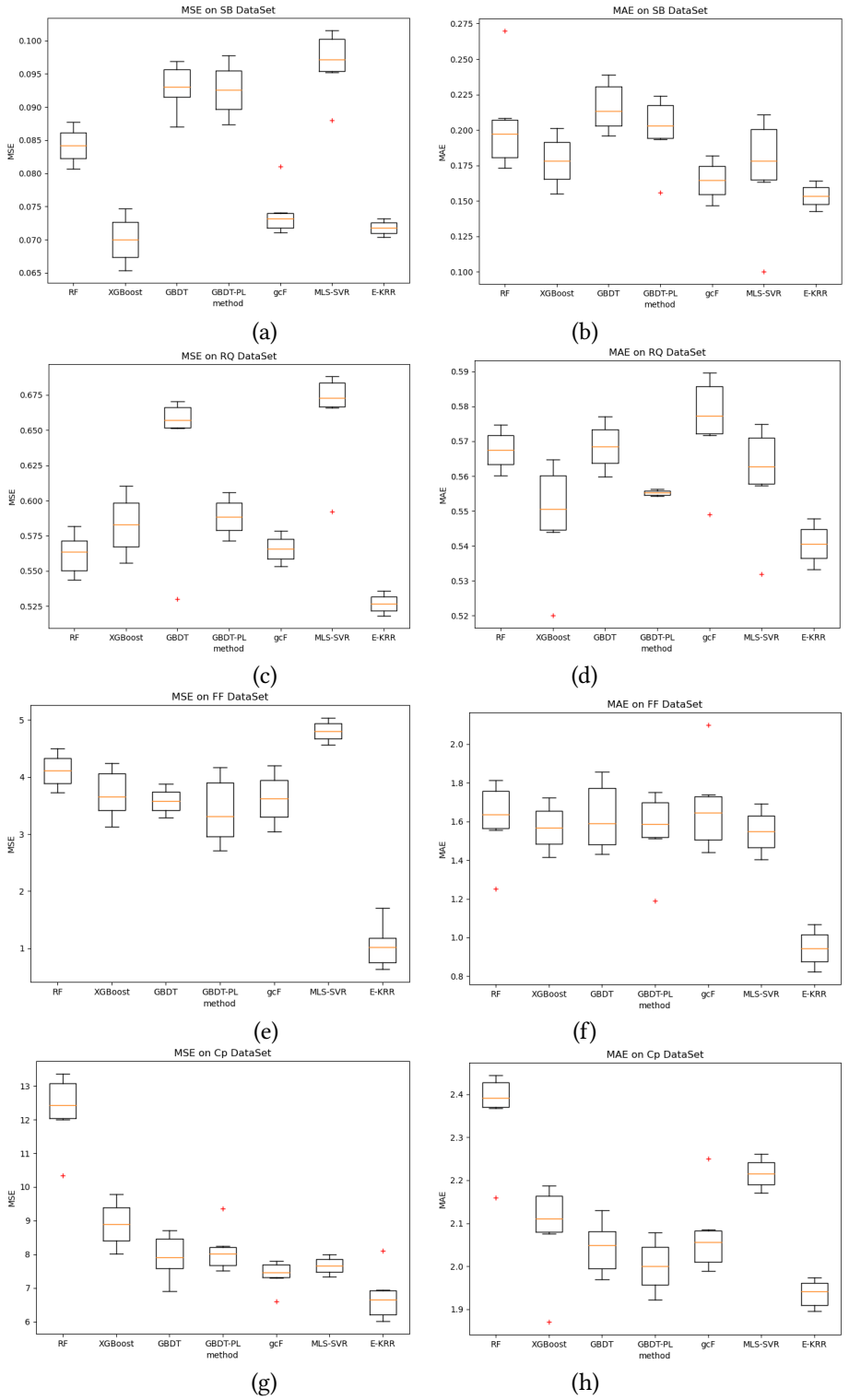


Fig. 2. The box plots of the various methods under MSE and MAE on UCI datasets.

Table 2. The average MSE on test sets of eight UCI datasets.

Datasets	RF	XGBoost	GBDT	GBDT-PL	gcF	MLS-SVR	**E-KRR**
SB	0.0842 ±0.0035	0.0800 ±0.0047	0.0941 ±0.0028	0.0926 ±0.0052	0.0726 ±0.0015	0.0984 ±0.0032	**0.0718** ±0.0014
Bi	5.7169 ±0.0292	4.8734 ±0.0731	3.8669 ±0.0462	3.7500 ±0.0774	3.5547 ±0.0422	4.8873 ±0.0757	**2.3041** ±0.0578
RQ	0.5582 ±0.0143	0.5829 ±0.0273	0.6609 ±0.0096	0.5886 ±0.0173	0.5658 ±0.0126	0.6772 ±0.0112	**0.5268** ±0.0089
Mg	0.0223 ±0.0016	0.0178 ±0.0495	0.0210 ±0.0340	0.0158 ±0.0573	0.0197 ±0.0154	0.0175 ±0.0319	**0.0152** ±0.0651
FF	4.1118 ±0.3876	3.8173 ±0.4283	3.5791 ±0.2945	3.5435 ±0.6284	3.6215 ±0.5755	4.8019 ±0.2384	**0.9142** ±0.2857
He	0.5579 ±0.0307	0.4931 ±0.0537	0.5207 ±0.0271	0.4944 ±0.0240	0.4990 ±0.0023	0.4833 ±0.0656	**0.4825** ±0.0545
Cp	12.686 ±0.6824	8.8966 ±0.8832	8.1312 ±0.5771	7.8751 ±0.3640	7.5471 ±0.2499	7.6637 ±0.3365	**6.4739** ±0.4710
Sp	0.0198 ±0.0528	0.0129 ±0.0216	0.0174 ±0.0721	0.0105 ±0.0630	0.0118 ±0.0008	0.0154 ±0.0219	**0.0093** ±0.0264

Table 3. The average MAE on test sets of eight UCI datasets.

Datasets	RF	XGBoost	GBDT	GBDT-PL	gcF	MLS-SVR	**E-KRR**
SB	0.1907 ±0.0175	0.1782 ±0.0231	0.2203 ±0.0184	0.2087 ±0.0154	0.1645 ±0.0176	0.1871 ±0.0237	**0.1535** ±0.0107
Bi	1.0822 ±0.0040	0.8942 ±0.0298	1.1417 ±0.0398	0.9969 ±0.0651	0.9568 ±0.0302	1.0925 ±0.0002	**0.8042** ±0.0762
RQ	0.5675 ±0.0073	0.5544 ±0.0104	0.5685 ±0.0086	0.5552 ±0.0010	0.5807 ±0.0091	0.5661 ±0.0088	**0.5406** ±0.0073
Mg	0.0946 ±0.0786	0.0998 ±0.0134	0.0911 ±0.0748	0.0876 ±0.0376	0.0862 ±0.0623	0.0858 ±0.0483	**0.0807** ±0.0400
FF	1.6846 ±0.1284	1.5694 ±0.1535	1.6628 ±0.1947	1.6313 ±0.1201	1.5901 ±0.1482	1.5481 ±0.1443	**0.9449** ±0.1232
He	0.7097 ±0.0213	0.5114 ±0.0169	0.5207 ±0.0214	0.5134 ±0.0359	0.5196 ±0.0582	0.5518 ±0.0797	**0.5025** ±0.0549
Cp	2.4061 ±0.0387	2.1320 ±0.0561	2.0275 ±0.0578	2.0006 ±0.0782	2.0373 ±0.0486	2.2158 ±0.0450	**1.9294** ±0.0343
Sp	0.0737 ±0.0484	0.0804 ±0.0406	0.0850 ±0.0038	0.0823 ±0.0540	0.0789 ±0.0028	0.0752 ±0.0463	**0.0704** ±0.0702

(RF) respectively. Based on the above discussion, we can conclude that our proposed method has better performance than other methods.

From Table 3, which presents the experimental results of the Mean Absolute Error (MAE), we can observe that GBDT and GBDT-PL methods perform poorly on SB and Sp datasets, which are 0.0668, 0.0146, 0.0552, 0.0119 higher than the optimal method (E-KRR) respectively. gcF method performs worst on the RQ dataset, which is 6.9% higher than the optimal method (E-KRR). However, GBDT-PL and gcF performed well on other datasets. Such as, on the Mg dataset, GBDT-PL and gcF

Table 4. Descriptive information for UCI datasets.

Datasets	Samples	Features	classes
CTG10	2126	21	10
MFCCs	7195	22	8
GenGap	4746	17	3
Obesity	2111	16	7
EGS	10000	14	2
Handwritten	5620	63	10

Table 5. The classification accuracy (%) on UCI datasets.

Datasets	RF	XGBoost	GBDT	MLS-SVR	KOC+	MENLR	**E-KRR**
CTG10	79.15	80.09	81.50	88.09	87.28	89.34	**91.38**
MFCCs	71.56	79.26	81.05	86.33	89.32	88.28	**91.61**
GenGap	81.80	81.46	81.88	88.63	89.64	83.15	**91.06**
Obesity	93.75	95.64	96.59	97.15	97.34	96.96	**97.72**
EGS	98.96	98.51	98.77	99.10	99.21	99.15	**99.47**
Handwritten	97.58	97.50	97.43	97.86	98.36	98.14	**98.47**

performed better, which are 12.2% and 13.6% lower than the worst method (XGBoost), respectively. At the same time, our method remains optimal in all datasets. In short, our regression method has strong robustness in terms of overall performance.

To further illustrate the advantages of our E-KRR method, we show the box diagram of MSE and MAE in Figure 2. The left and the right columns are MSE and MAE results respectively for SB, RQ, FF and Cp datasets.

Firstly, it can be seen from Figure 2 that among all regression methods, the MSE and MAE results of E-KRR are small, which means that the performance of our method E-KRR is better in most datasets. Meanwhile, the minimum MSE / MAE value of E-KRR is much smaller than that of other comparison methods. Secondly, from the perspective of data, the MSE and MAE results of the RF method are the largest among the seven regression methods on FF and Cp datasets. However, the RF method performs well on SB and Bi datasets. This shows that the RF method is unstable and the general degree is low. Similar to RF methods, gcF, GBDT-PL and MLS-SVR methods also have such problems. Therefore, it can be concluded that our method has good performance on UCI regression dataset, which helps us to find appropriate kernel and it's parameters in multi-kernel space.

4.3 Classification on UCI and image datasets

We also apply our E-KRR method to the dataset of classification task to verify the performance of our model. There are six comparison algorithms, including RF [37], XGBoost [10], GBDT [33], KOC+ [9], MENLR [43] and MLS-SVR [45], which are used to compare with our proposed method. Table 4 describes the detailed information of datasets used in the classification task, which includes Cardiotocography-10 (CTG10), Anuran Calls (MFCCs), GenderGap (GenGap), Obesity, Electrical Grid Stability (EGS), and Handwritten.

Table 5 shows the classification accuracy of our proposed method E-KRR and six comparison algorithms for the classification task. From the accuracy results, it can be seen that in the six classification datasets, the accuracy of the RF comparison algorithm is the lowest, which means its performance is the worst. On the Obesity dataset, the RF method is 9.26% lower than the optimal

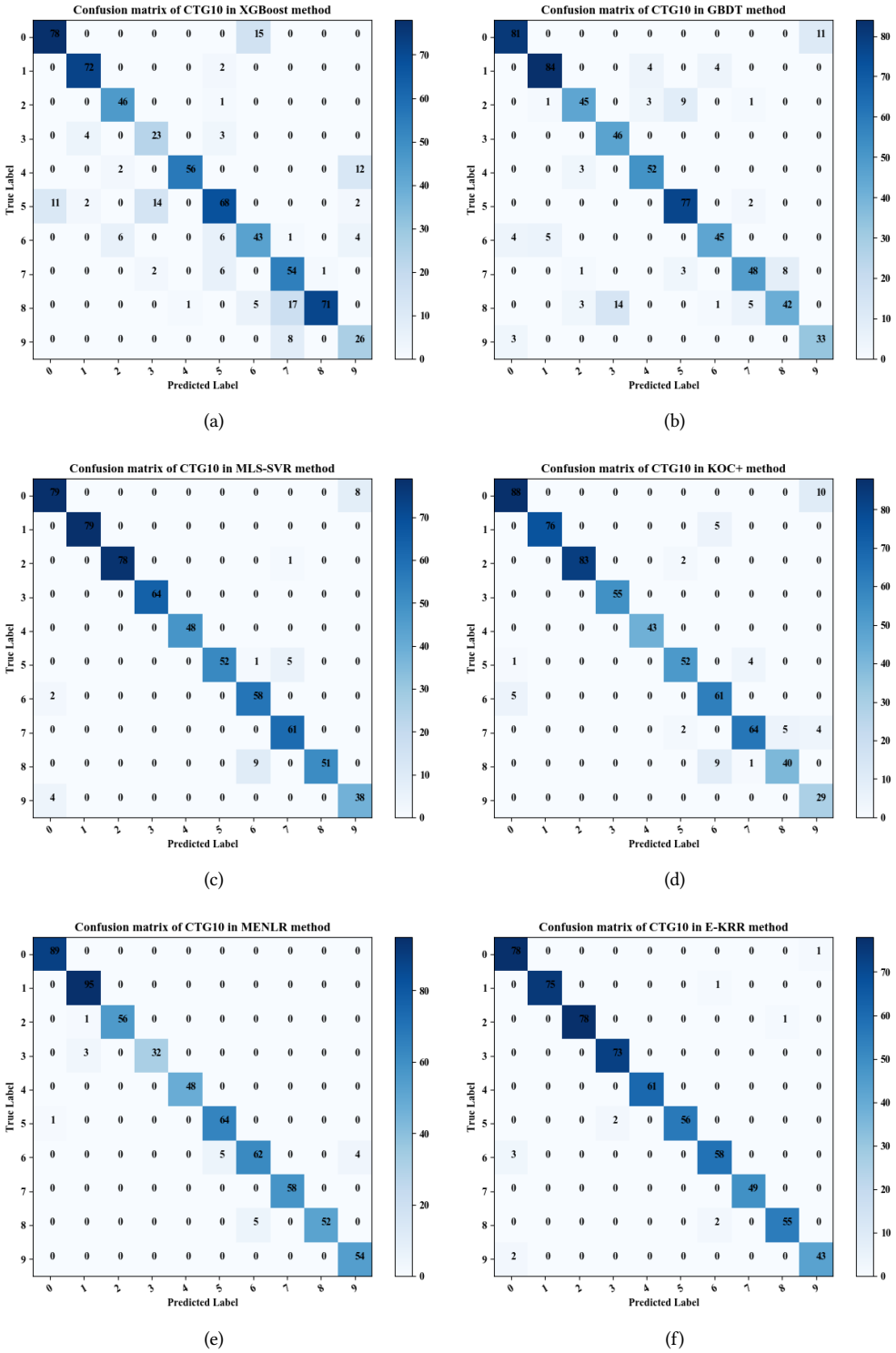
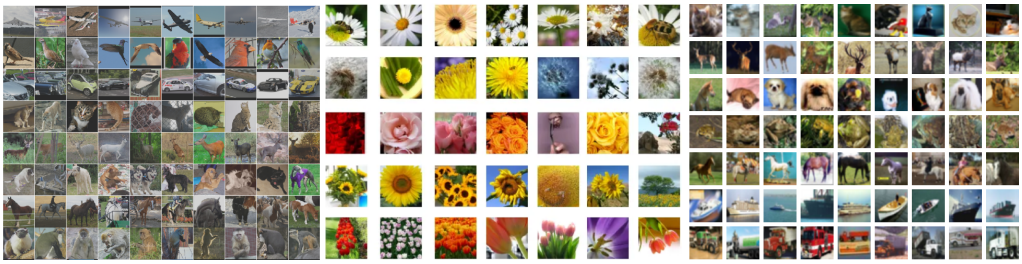


Fig. 3. Confusion matrix of CTG10 dataset on six methods.



(a) (b) (c)

Fig. 4. Sample images of the natural image datasets.

Table 6. The classification accuracy (%) on four natural image datasets.

Datasets	RF	XGBoost	GBDT	MLS-SVR	KOC+	MENLR	**E-KRR**
STL-10	37.04	53.96	47.02	75.06	71.01	76.51	**79.34**
CIFAR-10	57.88	79.35	82.14	91.70	91.62	92.50	**92.91**
Flower-17	77.22	83.76	82.47	88.95	90.63	92.09	**93.20**
ImageSeg	98.09	96.88	96.53	98.42	97.79	98.53	**98.83**

method (E-KRR) and 3.59% lower than the second optimal method (KOC+). Although the MENLR method performs well on CTG10 and MFCCs datasets, it performs poorly on GenGap datasets, which is 7.91% lower than our E-KRR method. On the other hand, on EGS and Handwritten datasets, all methods perform well. Among them, our E-KRR method performs best, which is 0.26% (KOC +) and 0.11% (KOC +) higher than the second best method, and 0.96% (XGBoost) and 1.04% (GBDT) higher than the worst method, respectively. Therefore, it can be concluded that our E-KRR method has good performance.

In order to more specifically demonstrate the advantages of our proposed method, figure 3 shows the confusion matrix of the CTG10 dataset on six methods, where the subgraph 3(a), 3(b), 3(c), 3(d), 3(e) and 3(f) represents XGBoost, GBDT, MLS-SVR, KOC+, MENLR and E-KRR methods respectively. From the subgraph 3(a), 3(b), 3(c) and 3(d), it can be seen that on the CTG10 dataset, the former four methods perform poorly, and there are many light shadow areas with misjudgment. Relatively speaking, the latter two methods perform better, where we can hardly see the light shadow area of misjudgment on the confusion matrix subgraph 3(f) of our E-KRR method, and the shadow of the correct judgment area on the diagonal is also deep. The above discussion shows that our method also maintains high accuracy and good performance in classification tasks.

We conducted experiments using six comparison algorithms on four natural image datasets, including STL-10, CIFAR-10, Flower-17 and Image-Segmentation (ImageSeg).

First, the STL-10 dataset is an image dataset that contains 10 types of objects, including airplanes, birds, cars, cats, deers, dogs, horses, monkeys, ships, and trucks. It is inspired by the CIFAR-10 dataset but with some modifications. In particular, each class has fewer labeled training examples than in CIFAR-10, but a very large set of unlabeled examples is provided to learn image models prior to supervise training. There are 1300 pictures in each category, of which 500 pictures are used for training and 800 pictures for testing. All examples are 96x96 pixel color images.

Second, the CIFAR-10 dataset is an RGB color picture with 10 categories, including airplanes, automobiles, birds, cars, deer, dogs, frogs, horses, ships, and trucks. The size of the pictures are 32×32. There are 50000 training pictures and 10000 test pictures in the dataset.

Third, the Flower-17 dataset contains a total number of 1360 images with 17 species of flowers and 80 images per category, which is selected by the visual geometry group of Oxford University.

Fourth, the ImageSeg dataset is randomly selected from a database containing seven outdoor images, including brick face, sky, foliage, cement, window, path and grass. These images are manually segmented to create a classification for each pixel.

Some selected images of STL-10, Flower-17 and CIFAR-10 datasets are shown in Figure 4 below. Table 6 shows the accuracy of our method E-KRR and six comparison algorithms on natural image data sets.

The experimental results in Table 6 show that our method E-KRR has the highest experimental accuracy on natural image data sets. It can be noted that each method performs poorly on the STL-10 dataset. However, our method E-KRR is 2.83% higher than the second best method (MENLR) and 42.3% higher than the worst method (RF). RF method performs poorly on STL-10, CIFAR-10 and Flower-17 data sets. However, the RF method performs better on the Imageseg data set, but it is still 0.74% lower than the best method (E-KRR).

In order to clearly and intuitively show the advantages of our proposed method E-KRR in the comparison algorithm, we further explain the experimental results in figure 5. From figure 5, we can clearly see that the histogram of E-KRR is higher than that of other comparison methods, which means that our proposed model can select the best kernel combination and its parameters in a diversified multi-solution space. Therefore, compared with other methods, our method E-KRR has better performance and can effectively improve the classification accuracy of natural image data set recognition.

Therefore, we can conclude that the proposed E-KRR model not only has good performance in regression tasks, but also maintains high accuracy in classification and natural image recognition tasks. Therefore, it fully proves that our proposed E-KRR method can select the appropriate kernel function and its parameters in multiple Reproducing Kernel Hilbert Spaces.

5 CONCLUSION

In this paper, to solve the selection problem of kernel function and its parameters in multi-kernel space an ensemble kernel ridge regression model (E-KRR) is proposed. E-KRR assumes original data have multi-view presentations and therefore transforms multi-view representation into multi-kernel representations.. Unlike the traditional single KRR method, in the proposed E-KRR method finds the best combinational kernel representations and their corresponding kernel parameters in multiple Reproducing Kernel Hilbert Spaces (RKHSs). Therefore, E-KRR obtains the ensemble multiple kernel space that enhances the performance of the model. Regression and classification experiments on UCI and image datasets show that our proposed method has the lowest loss and the highest accuracy compared with the state-of-art methods.

In future work, we will study the relationship between multi-kernel representation and deep learning while considering the combination of multi-kernel representation and deep neural network. The kernel function will be added to the network of each layer for reverse transmission by appropriately increase the width and depth of the network.

6 ACKNOWLEDGMENTS

This research was funded in part by Primary Research & Development Plan of Jiangsu Province (BE2018627).

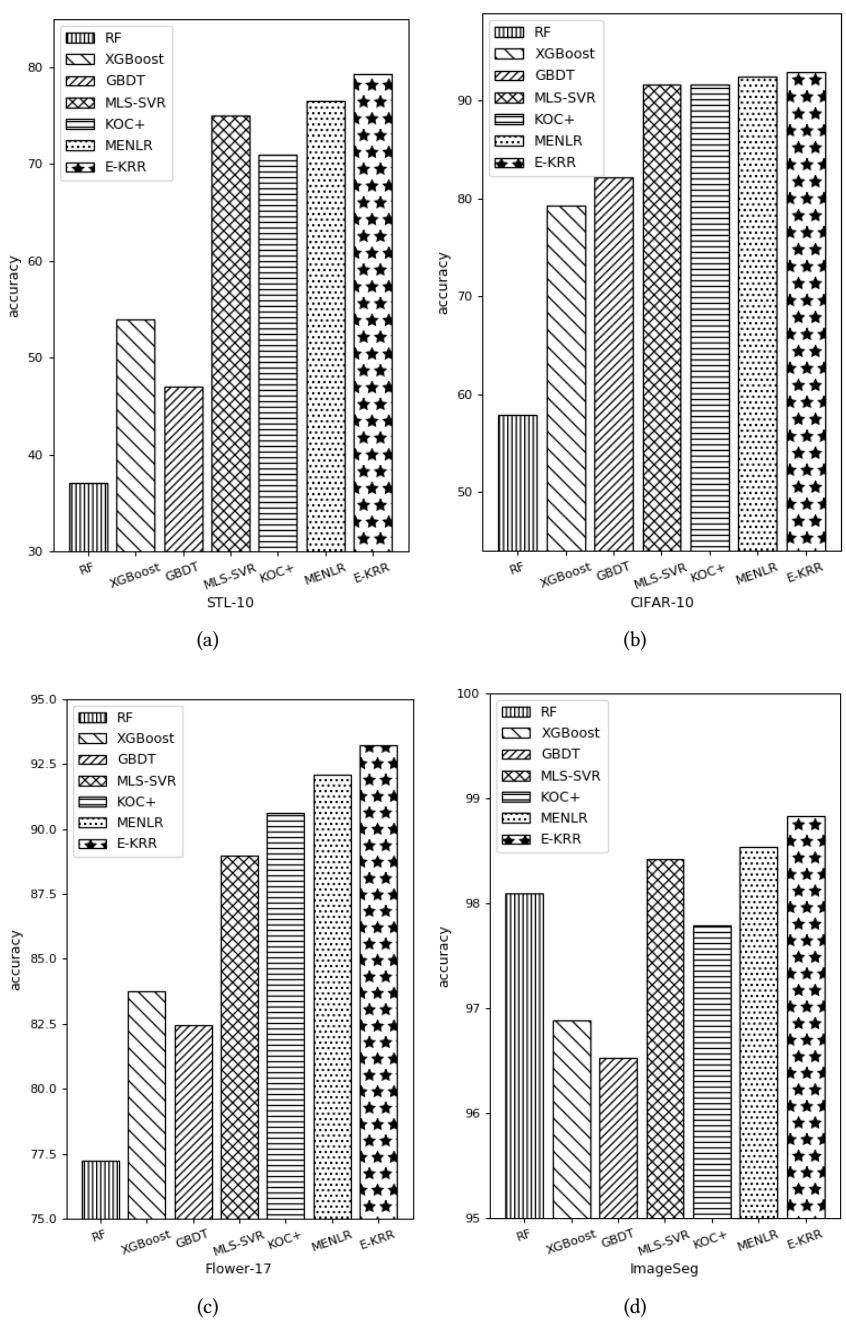


Fig. 5. The classification accuracies (%) of seven methods on natural image datasets .

REFERENCES

[1] 2016. Learning multi-kernel multi-view canonical correlations for image recognition. *Computational Visual Media* 2 (2016), 10.

[2] Ahmad and Ghodselahi. 2011. A Hybrid Support Vector Machine Model for Credit Scoring. *international journal of computer applications* 17, 5 (2011), 1–5.

[3] R. Alejo, V García, AI Marqués, JS Sánchez, and JA Antonio-Velázquez. 2013. Making Accurate Credit Risk Predictions with Cost-Sensitive MLP Neural Networks. (2013).

[4] Salah Althloothi, Mohammad H. Mahoor, Xiao Zhang, and Richard M. Voyles. 2014. Human activity recognition using multi-features and multiple kernel learning. *Pattern Recognition* 47, 5 (2014), 1800–1812.

[5] A. Atiya. 2005. Learning with kernels: Support vector machines, regularization, optimization, and beyond. *IEEE Transactions on Neural Networks* 16, 3 (2005).

[6] B. Baesens, T Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54, 6 (2003), 627–635.

[7] Vanya Van Belle and Paulo Lisboa. 2014. White box radial basis function classifiers with component selection for clinical prediction models. *Artificial Intelligence in Medicine* 60, 1 (2014), 53–64.

[8] R. Benedict, K. Nicolas, H Marius, N. K. Speicher, and P. Nico. 2019. web-rMKL: a web server for dimensionality reduction and sample clustering of multi-view data based on unsupervised multiple kernel learning. *Nucleic Acids Research* W1 (2019), W1.

[9] A Cg, A At, and B Mt. 2019. KOC+: Kernel ridge regression based one-class classification using privileged information. *Information Sciences* 504 (2019), 324–333.

[10] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *ACM* (2016).

[11] F. Douak, F. Melgani, and N. Benoudjit. 2013. Kernel ridge regression with active learning for wind speed prediction. *Applied Energy* 103, MAR. (2013), 328–340.

[12]]FengCascading Z. A. Feng, C Qzb, E Dsd, and J. F. Liu. [n. d.]. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Applied Soft Computing* 84 ([n. d.]).

[13] C. Gautam, A. Tiwari, and M. Tanveer. 2020. AEKOC+: Kernel Ridge Regression-Based Auto-Encoder for One-Class Classification Using Privileged Information. *Cognitive Computation* 12, 2 (2020), 412–425.

[14] F. Hansen and G. K. Pedersen. 2003. JENSEN'S OPERATOR INEQUALITY. *Bulletin of the London Mathematical Society* (2003).

[15] J. He, L. Ding, J. Lei, and M. Ling. 2014. Kernel ridge regression classification. *IEEE* (2014).

[16] S. Li, Y. Lin, T. Zhu, M. Fan, and S. Xu. 2021. Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method. *Neural Computing and Applications* 11 (2021).

[17] Z. Lin, M. Chen, and Y. Ma. 2010. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *eprint arxiv* 9 (2010).

[18] Stan Lipovetsky. 2006. Two-parameter ridge regression and its convergence to the eventual pairwise model. *Mathematical & Computer Modelling* 44, 3-4 (2006), 304–318.

[19] S. Mukkamala, Hsa Andrew, and A. Abraham. 2003. Intrusion Detection Using Ensemble of Soft Computing Paradigms. *Springer Berlin Heidelberg* (2003).

[20] J. Naik, P. Satapathy, and P. K. Dash. 2017. Short-Term Wind Speed and Wind Power Prediction using Hybrid Empirical Mode Decomposition and Kernel Ridge Regression. *Applied Soft Computing* (2017), 1167–1188.

[21] L. Nanni and A. Lumini. 2009. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications* 36, 2-part-P2 (2009), 3028–3033.

[22] K. Rakesh and P. N. Suganthan. 2017. An Ensemble of Kernel Ridge Regression for Multi-class Classification. *Procedia Computer Science* 108 (2017), 375–383.

[23] M. Salhov, O. Lindenbaum, Yariv Aizenbud, A. Silberschatz, Y. Shkolnisky, and A. Averbuch. 2016. Multi-View Kernel Consensus For Data Analysis. (2016).

[24] Hafizi Abu Samah, Nor Ashidi Mat Isa, and Kenny Kal Vin Toh. 2015. Automatic false edge elimination using locally adaptive regression kernel. *Signal,Image&Video Processing* (2015).

[25] C. Saunders, A. Gammerman, and V. Vovk. 1998. Ridge Regression Learning Algorithm in Dual Variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998.*

[26] A Sc, B Aaa, and A Jpt. 2005. Feature deduction and ensemble design of intrusion detection systems. *Computers & Security* 24, 4 (2005), 295–307.

[27] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. 2012. Facial Action Recognition Combining Heterogeneous Features via Multi-Kernel Learning. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics* 42, 4 (2012), 993–1005.

[28] Y. Shi, J. Li, and Z. Li. 2018. Gradient Boosting With Piece-Wise Linear Regression Trees. (2018).

[29] Yu Shi, Jian Li, and Zhize Li. 2019. Gradient Boosting with Piece-Wise Linear Regression Trees. In *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*.

[30] H. K. Sun, D. H. Cho, and K. H. Seok. 2012. Study on the ensemble methods with kernel ridge regression. *Journal of the Korean Institute of Information & Communication Engineering* 23, 2 (2012).

[31] Jun Sun, Zhiyong Shen, Hui Li, and Yidong Shen. 2008. Clustering Via Local Regression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[32] J. T. Tsai, Y. Y. Lin, and H. Y. M. Liao. 2014. Per-Cluster Ensemble Kernel Learning for Multi-Modal Image Clustering With Group-Dependent Feature Selection. *IEEE Transactions on Multimedia* 16, 8 (2014), 2229–2241.

[33] Wang, YZ, Feng, DW, Ii, DS, Chen, XY, Zhac, and YX. 2016. A mobile recommendation system based on Logistic Regression and Gradient Boosting Decision Trees. *IEEE Ijcnn* (2016).

[34] H. Wang, Q. Xu, and L. Zhou. 2015. Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble. *Plos One* 10 (2015).

[35] Kilian Q. Weinberger and Gerald Tesauro. 2007. Metric Learning for Kernel Regression. *Journal of Machine Learning Research* 2 (2007), 612–619.

[36] Wu and Hao. 2019. Research on Diabetes Prediction Model Based on XGBoost Algorithm. *International Conference on Advanced Materials and Computer Science (ICAMCS) 2019*.

[37] Hongyan Wu, Yunpeng Cai, Yongsheng Wu, Ren Zhong, Qi Li, Jing Zheng, Denan Lin, and Ye Li. 2017. Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression. *BioScience Trends* 11, 3 (2017), 292.

[38] Z. Xiao, M. H. Mahoor, and R. M. Voyles. 2013. Facial expression recognition using HessianMKL based multiclass-SVM. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*.

[39] Jian Rong Yao and Jia Rui Chen. 2019. A New Hybrid Support Vector Machine Ensemble Classification Model for Credit Scoring. *Journal of Information Technology Research* 12, 1 (2019), 77–88.

[40] Y. Ye, X. Liu, J. Yin, and E. Zhu. 2016. Co-regularized kernel k-means for multi-view clustering. In *2016 23rd International Conference on Pattern Recognition (ICPR)*.

[41] L. Zhang and P. N. Suganthan. 2017. Benchmarking Ensemble Classifiers with Novel Co-Trained Kernel Ridge Regression and Random Vector Functional Link Ensembles [Research Frontier]. *IEEE Computational Intelligence Magazine* 12, 4 (2017), 61–72.

[42] Null Xiao Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn. 2014. A lp-norm MTMKL framework for simultaneous detection of multiple facial action units. In *IEEE Winter Conference on Applications of Computer Vision*.

[43] Z. Zheng, Z. Lai, X. Yong, S. Ling, W. Jian, and G. S. Xie. 2017. Discriminative Elastic-Net Regularized Linear Regression. *IEEE Transactions on Image Processing* (2017).

[44] Zhi Hua Zhou and Ji Feng. 2017. Deep Forest: Towards An Alternative to Deep Neural Networks. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*.

[45] X. Zhu and Z. Gao. 2018. An efficient gradient-based model selection algorithm for multi-output least-squares support vector regression machines. *Pattern Recognition Letters* 111, AUG.1 (2018), 16–22.