A Time Series Data Regression Model Based on Multi-kernel KPCA Dimension Reduction and Bagging Algorithms

Sitian Xu [1,*], Wenfeng Xu [2]

[1] School of Mathematics, Physics and Civil Engineering, Beijing Institute of Technology, Zhuhai, Zhuhai Guangdong Province, 519088

[2] School of Accounting and Finance, Beijing Institute of Technology, Zhuhai, Zhuhai, Guangdong Province 519088

* Corresponding Author Email: vanessa07202003@163.com

Abstract. Regression models for high-dimensional data have always been a hot topic in the field of statistical learning. Considering the case that the predictor variable is a high-frequency time series and the response variable is a continuous scalar, this paper proposes a regression method based on a multi-kernel KPCA Dimension reduction method and the Bagging algorithms. The proposed method adaptively solves the problem of kernel function selection and unsupervised ness in KPCA Dimension reduction by splicing the projection data under various kernel functions and using the Bagging algorithms to mine the relationship between projection data and the response variable. In the real data analysis, this paper selects the multiple regression model, the LASSO regression based on model selection, the multiple regression model based on PCA Dimension reduction and single-kernel KPCA Dimension reduction as comparison methods. The results show that the proposed method has better performance than other comparison methods. Since the basic regressor in the Bagging framework is model-free, some prediction models can be used to improve the prediction accuracy for more complex data situations.

Keywords: KPCA, Bagging algorithms, time series data, high-dimensional regression.

1. Introduction

Time-series data is a common form of data that has occupied a large amount of storage space in fields such as commerce, industry, medicine, and scientific research. In recent years, time series data regression has been widely used in many fields, such as environmental epidemiology, infectious disease prediction and power resources. However, unlike traditional multivariate data, due to the characteristics of time-series data such as high Dimension, large quantity, many features, and data information changing with time, the prediction task for time-series data faces great challenges in practical applications. This paper considers the case where the predictor variable is a high-frequency time series and the response variable is a continuous scalar; taking the meat data set and corn data set as examples. In these two data sets, the predictor variable is a high-frequency time series measured by infrared spectrometer, and the response variables are water content, protein content, fat content, oil content and starch content. The specific data task is to predict the content of these nutrients based on high-frequency time series data.

Aiming at the time series data prediction problem of appeal, the simplest method is usually to regard it as multivariate data and then study the regression problem between a dependent variable and multiple independent variables [1]. Because of its practicability and effectiveness, multiple data regression methods are widely used in more and more fields in today's society. Multiple regression analysis is a statistical method to establish a forecasting formula (empirical formula) based on statistical data [2]. For example, Zhang Jie [3] et al. predicted the total amount of garbage based on multiple regression Li Jinhai et al. [4] put forward the application steps of the multiple regression method based on the multiple regression mathematical model. Wang Quanzhong et al [5] introduced two types of multiple regression models, but the multiple regression models are prone to fall into curse of Dimension when the predictor is high dimension. Time series data regression is a typical high-dimensional problem. In this case, we generally use Dimension reduction technology to deal

with high-dimensional problems first and then use the regression model to predict the response variable, and the most typical Dimension reduction technique is PCA principal component analysis Dimension reduction, this method retains the components with large variance and more information through linear transformation and discards the direction with less information, thereby reducing the Dimension of the data and removing the correlation between the original variables. Yue Fengfeng [6] et al. conducted principal component analysis and established multiple regression prediction for peony flowering period. Ji Li [7] et al. statistically analyzed the characteristics of the early flowering period from 2007 to 2021, and used the PCA principal component analysis method to predict the early flowering period of Wintersweet. However, there are two problems in the original PCA: first, this method can only deal with linear dimension reduction, and the chain structure between each principal component and the response variable is unknown, that is, unsupervised.

Aiming at the former, some improved Dimension reduction methods have emerged, the most typical of which is the Kernel Principal Component Analysis (KPCA) proposed by Li Zhiqin et al. [8], which combines low-dimensional. The linearly inseparable data in the space is mapped to the high-dimensional space, making the data linearly separable in the high-dimensional feature space. Yang Xuexue et al. [9] used nuclear principal component analysis to reduce the dimension and optimize the data of the hazard factors affecting storm surge disasters, and then output the predicted value. Fu Hua et al. [10] used KPCA to predict gas emission. For high-dimensional data, model selection [11] is generally used as a processing method. However, model selection methods ignore additional uncertainties, often biasing the predictive model and underestimating the variance [12]. In addition, different selection methods or criteria may produce different optimal models.

Based on this, in the context of regression prediction of time series data, in order to solve the kernel function selection and unsupervised problems of KPCA, this paper proposes a regression model based on multi-kernel KPCA and Bagging algorithm. The innovations of the proposed method are mainly as follows: The proposed method can adaptively solve the selection problem of the kernel function in KPCA dimensionality reduction by splicing projection matrices under various kernel functions; Using the Bagging algorithms in multi-model fusion The relationship between each principal component and the response variable can be mined under the following conditions, and at the same time, the variance and bias of the prediction model can be effectively weighed. The actual data analysis shows that the proposed method has certain competitiveness compared with the comparison method.

2. Theory and Method

2.1. Kernelized Principal Component Analysis

In order to compensate for the lack of linear dimensionality reduction capability, nonlinear dimensionality reduction methods based on kernel functions are proposed, and the more commonly used method is Kernelized Principal Component Analysis (KPCA); that is, an algorithm that converts data from low-dimensional nonlinear to high-dimensional linearly separable data. Specifically, suppose there is an original dataset $D=\{x_1, x_2..., x_m\}$, where $xi =(x_{i1}, x_{i2,}..., x_{ip})'$, m is the sample size, p is the data dimension, and the mapping of x_i in the high-dimensional feature space is $z_i = \phi(x_i)$, the data set in the high-dimensional feature space is $D'=\{z_1, z_2,..., z_m\}$, and its covariance matrix is:

$$\Sigma = \sum_{i=1}^{m} z_i z_i^T \tag{1}$$

The solution objective of kernel principal component analysis is:

$$\begin{cases} z_j = \omega_j^T \phi(x) \\ \left(\sum_{i=1}^{m} z_i z_i^T \right) \omega_j = \lambda_j \omega_j, \; j = 1, 2, \ldots, m \end{cases}$$

Among them, ω_j is the eigenvector corresponding to the eigenvalue λ_j of Σ, and z_j is the jth coordinate component after the sample projection. Introduce kernel function:

$$\kappa\left(x_i, x_j\right) = \phi^T\left(x_i\right)\phi\left(x_j\right) \tag{2}$$

Let $K = \left(\kappa\left(x_i, x_j\right)\right)_{m\times m}$ be the kernel matrix corresponding to K (x_i, x_j) $\alpha^j = \left(\alpha_1^j, \alpha_2^j, \ldots, \alpha_m^j\right)^T$ and be the eigenvector corresponding to the jth largest eigenvalue λ_j of the kernel matrix K. The jth coordinate component after sample:

$$z_j = \omega_j^T\phi(x) = \sum_i^m \alpha_i^j \phi^T\left(x_i\right)\phi(x) = \sum_i^m \alpha_i^j \kappa\left(x_i, x\right) \tag{3}$$

Where α_j is the normalized vector. Table 1 lists several commonly used kernel functions:

Table 1. Commonly used kernel functions and related descriptions

Name	expression	parameter
linear	$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$	
polynomial	$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\right)^d$	d>1 is the degree of the polynomial
Gaussian	$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right)$	$\sigma > 0$ is the bandwidth of the Gaussian kernel (width)
Sigmoid	$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \tanh\left(\beta\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \theta\right)$	$\beta > 0, \theta < 0$

2.2. Bagging algorithms

Bagging algorithms is an integrated learning algorithms in the field of machine learning. Originally proposed by Leo Breiman [13] in 1996. The main process of Bagging method is to randomly sample T training data sets from the initial training set with replacement, and use the sampled training set samples to train the base learner. Specifically, by constructing multiple weak learners separately, the multiple weak learners are parallel to each other and can be trained at the same time. Finally, multiple weak learners are combined to achieve the purpose of balancing deviation and variance. The steps of the Bagging algorithms are shown in Figure 1 below:

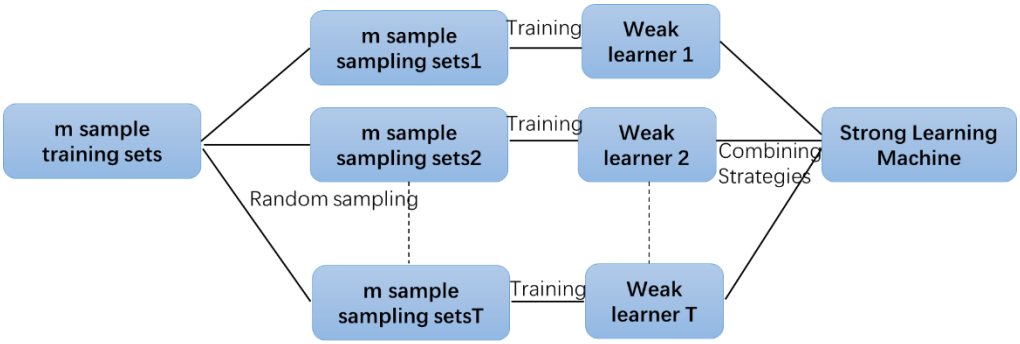


Figure 1. Introduction to Bagging algorithms

2.3. High Dimensional Time Series Data Regression Model Based on Multi-kernel KPCA-Bagging

Considering the selection of the KPCA kernel function and how to construct the relationship between each principal component and response variable, this paper uses the idea of model averaging instead of model selection to adaptively solve the selection of kernel function in KPCA Dimension reduction, and based on the Bagging algorithms to The relationship between each principal

component vector and the response variable in different regenerated kernel Hilbert spaces is captures to obtain the final combined prediction output. The basic flow of the algorithms is as follows:

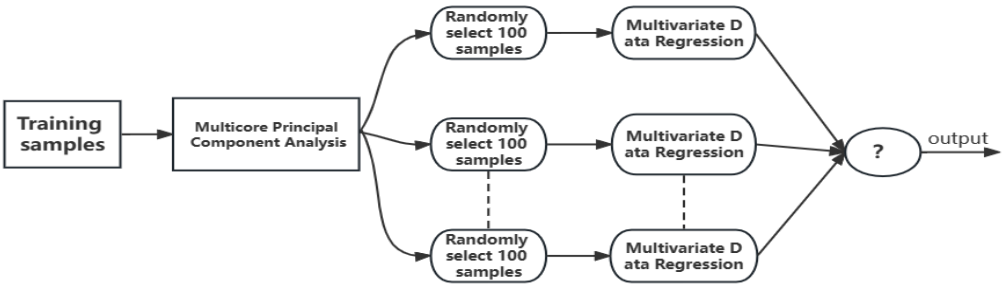


Figure 2. Multi-kernel KPCA—Bagging algorithms flow

Step1. Divide the data set into a training set and a test set;

Step2. On the training set, according to different kernel functions, the corresponding kernel function matrix is calculated for eigenvalue decomposition, and the corresponding projection data is obtained according to the cumulative contribution rate and spliced to form the feature matrix on the training set and test set respectively;

Step3. Use the feature matrix on the training set to estimate the parameters of each sub-model in the Bagging model, and bring the feature matrix on the test set into the predicted value of the response variable on the test set. The specific flow chart is shown in Figure 2.

3. Data analysis

3.1. Data source and description

The experimental data is the spectral data set of meat samples downloaded from **http://lib.stat.cmu.edu/datasets/tecator** and **http://www.eigenvector.com/data/** [14]. The corn sample spectrum data set downloaded from Corn/index.html, in which there are 240 samples in the meat sample data set, each sample contains the water, fat and protein content of the meat sample, and the near-infrared spectroscopy analyzer in the band the absorption spectrum measured in the range of 850-1050 nanometers (nm). A total of 100 absorption spectrum data were recorded at intervals of every two bands, that is, the dimension is 100 dimensions. The data set of corn samples consists of n=80 corn samples, each sample contains the moisture, oil and protein content of corn, and the measured band range is 1100-2498 nanometers, with an interval of 2 nanometers, a total of 700 absorption spectrum data, that is, the dimension is 700 dimensions.

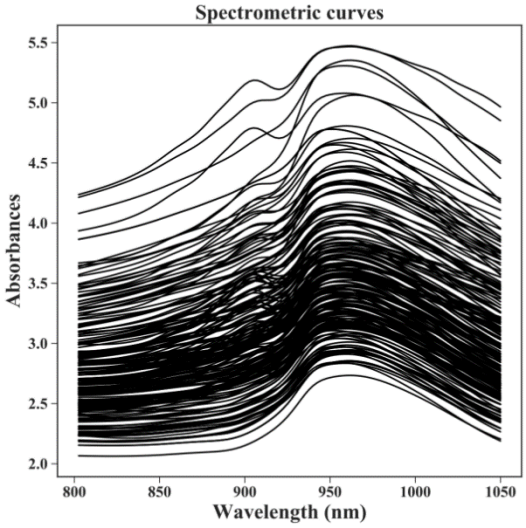


Figure 3. The wavelength of meat data set

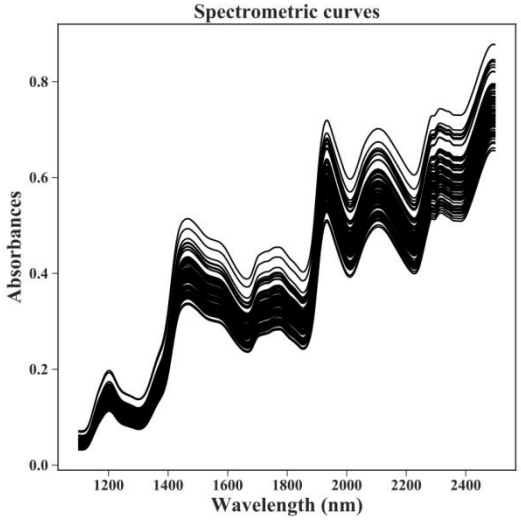


Figure 4. The wavelength of corn data set

The experiment is based on the sklearn framework in Python 3.7. The sklearn version is 0.23. In addition, data science libraries such as numpy, scipy, and matplotlib are also used. In order to compare the advantages and disadvantages of the methods, this paper uses a total of five comparison models, including multiple linear regression, linear regression based on PCA Dimension reduction, LASSO based on model selection, and linear regression based on radial basis function KPCA, to compare the meat data set and Maize dataset for regression prediction. And compare the performance of each model on the test set after training on the same training set. To evaluate the predictive performance of the model, this paper uses the commonly used evaluation indicators Mean Squared Error (Mean Squared Error, MSE) and Mean Relative Error (Mean Relative Error, MRE) in the regression prediction problem to evaluate the predictive performance of the model. The specific formula is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2 \quad MRE = \frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i - \hat{y}_i\right)\right| \tag{4}$$

3.2. Experimental results

Based on the above data sets and experimental evaluation indicators, a regression prediction experiment was designed, and the experiment was repeated 100 times, with the test set accounting for 30%. In this paper, six methods are used to calculate the mean and standard deviation of the MSE and MRE of the moisture, fat and protein content in the meat data set and the corn data set under three different frequency waves. As shown in Table.2-Table.7 and Figure.5-Figure.10. It can be seen from the above chart that compared with the other five models, the mean square error and absolute error of 100 experiments are relatively small for the proposed method.

Table 2. Comparison of fat content prediction results under the meat data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.651638	0.081539	0.947119	0.319724
LR	0.969450	0.224151	4.209465	5.131105
PCA_LR	1.089862	0.108381	2.590884	0.634465
TREE	1.370323	0.197731	4.226789	1.114247
LASSO	1.108447	0.109388	2.647127	0.627993
KPCA_LR	4.476377	0.398752	33.193343	5.997941

Table 3. Comparison of moisture content prediction results under the meat data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.613984	0.077861	0.876496	0.306082
LR	0.864306	0.175900	3.064568	2.906257
PCA_LR	1.113952	0.118217	2.552025	0.552218
TREE	1.189492	0.189511	3.381611	1.375295
LASSO	1.122767	0.115619	2.545209	0.544743
KPCA_LR	3.681273	0.324396	22.055013	3.889160

Table 4. Comparison of protein content prediction results under the meat data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.514019	0.055944	0.554844	0.167050
LR	0.801386	0.172937	2.349426	2.146189
PCA_LR	0.997046	0.122854	2.386091	0.740505
TREE	1.072769	0.168640	3.086963	1.075280
LASSO	1.017949	0.122738	2.416454	0.729215
KPCA_LR	1.501647	0.157330	4.057700	0.842568

Table 5. Comparison of prediction results of moisture content under corn data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.127808	0.020955	0.024586	0.007535
LR	0.151488	0.020325	0.035507	0.009607
PCA_LR	0.246579	0.031746	0.094093	0.022181
TREE	0.331166	0.055946	0.189773	0.061315
LASSO	0.314716	0.036111	0.149946	0.029429
KPCA_LR	0.338853	0.104771	0.354930	0.472704

Table 6. Comparison of starch content prediction results under corn data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.304099	0.044711	0.149542	0.046067
LR	0.372171	0.050988	0.222209	0.072827
PCA_LR	0.339368	0.040467	0.170557	0.039793
TREE	0.430520	0.067175	0.311485	0.090838
LASSO	0.663346	0.086392	0.680530	0.152587
KPCA_LR	0.412412	0.078577	0.293998	0.192349

Table 7. Comparison of oil content prediction results under corn data set

methods	MRE		MSE	
	Mean	Std	Mean	Std
MKPCAB	0.075687	0.011363	0.009792	0.002973
LR	0.099083	0.019357	0.018517	0.006965
PCA_LR	0.138513	0.015248	0.029405	0.006975
TREE	0.172771	0.025205	0.051444	0.013739
LASSO	0.140405	0.016255	0.031350	0.006006
KPCA_LR	0.187387	0.065930	0.098660	0.173548

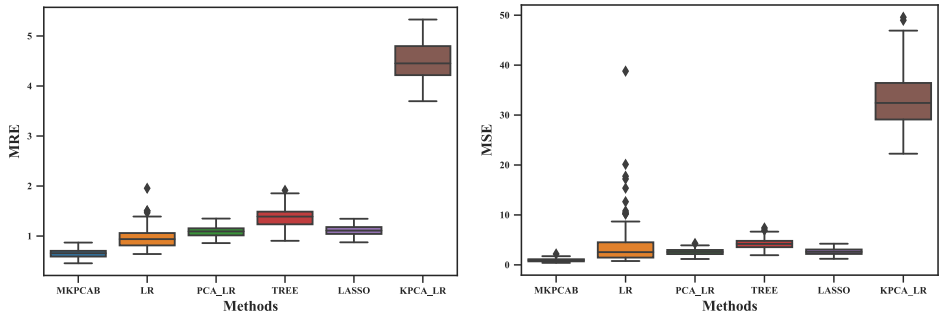


Figure 5. The distribution of fat content prediction error under the meat dataset

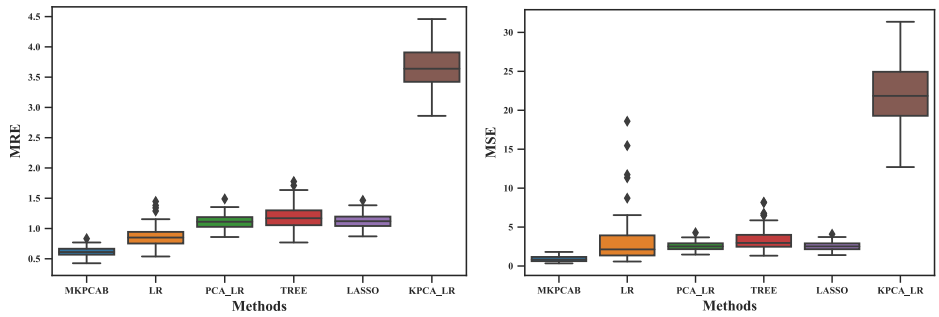


Figure 6. The distribution of moisture content prediction error under the meat data set

In addition, it can be clearly observed that compared with single-kernel KPCA, multi-kernel KPCA has shown a significant optimization effect. The main reason may be: compared with KPCA, the combination of multi-kernel KPCA and Bagging algorithms can adaptively select effective features, so that achieve higher accuracy. This can also be seen by observing the results of multiple linear regression. Specifically, there may be a linear relationship between the features at these time points and the response variable, but there is multicollinearity, so the performance of linear regression based on PCA is also relatively low. Well, the proposed method will also adaptively favour linear kernels. Therefore, the performance is not inferior to other methods.

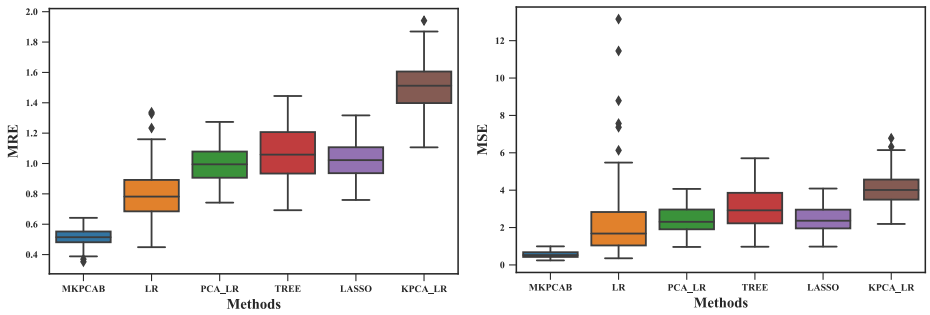


Figure 7. Protein Content Prediction Error Distribution for Meat Data set

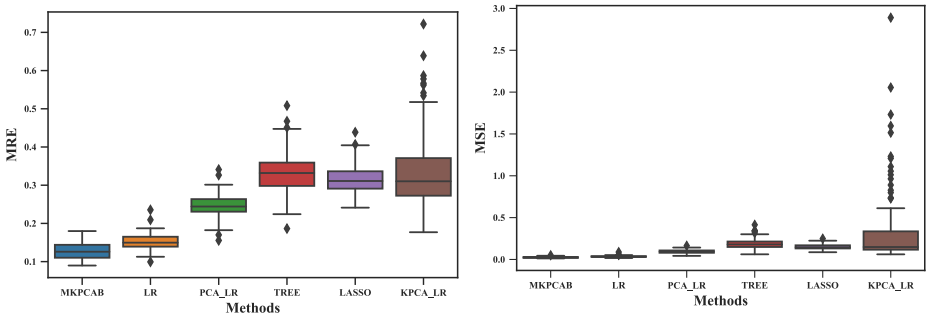


Figure 8. The distribution of moisture content prediction errors under the corn data set

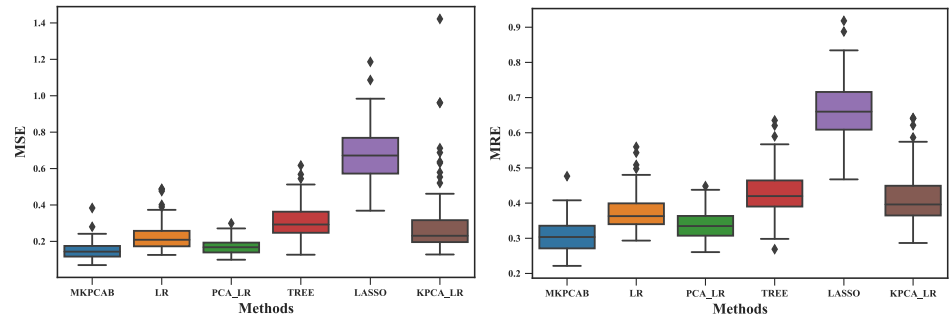


Figure 9. Starch content prediction error distribution under the corn data set

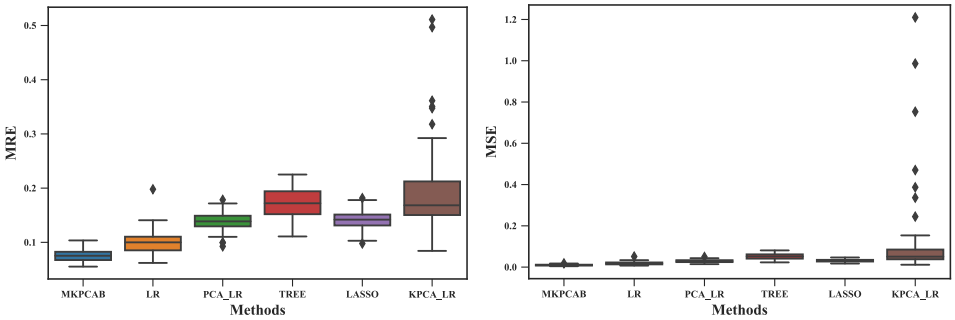


Figure 10. The distribution of oil content prediction errors under the corn data set

4. Conclusion and discussion

Time series data regression is widely used in many fields. In view of the high Dimension of variables, high redundancy, selection of kernel functions in KPCA and unsupervised characteristics of data sets, this paper proposes a multi-kernel KPCA and Bagging A high-dimensional time series data regression model combined with algorithmss. The proposed method uses the idea of model averaging instead of model selection, and adaptively solves the problem of kernel function selection in KPCA Dimension reduction; in addition, since KPCA is an unsupervised Dimension reduction method, the proposed method uses the Bagging algorithms to adaptively to capture the relationship between each principal component and the response variable in a nonlinear space, the actual data analysis results show that the proposed method has higher accuracy than other comparative methods.

For future research directions: On the one hand, the feasible way to improve multi-kernel KPCA is to construct a weighted KPCA kernel function matrix to improve the efficiency of data dimension reduction and the recognition performance of machine learning regression algorithmss, but the calculation efficiency of the kernel matrix is an unsolved problem. On the other hand, Bagging weighting is average weighting, which does not take into account that the accuracy and importance of each model are different. Based on this, we can use the weighted Bagging algorithms, such as the correlation coefficient weighting based on the heuristic model average, the information gain weighting [15].

References

[1] Luo Laihui. A Review of Tensor Regression Model and Its Application Research [J]. Statistics and Application, 2020, 9: 855.

[2] Keith T Z. Multiple regression and beyond: An introduction to multiple regression and structural equation modeling [M]. Routledge, 2019.

[3] Zhang Jie, Zhu Jiaming. Prediction of Total Garbage Based on Multiple Regression and Triple Exponential Smoothing Method [J]. Journal of Qilu University of Technology, 2020, 34(04): 69-74.

[4] Jinhai Li. Application of Multiple Regression Analysis in Forecasting [J]. Journal of Hebei University of Technology, 1996, (3): Volume 15.

[5] Quanzhong Wang. Logistic Regression Model for Two Types of Related Data Analysis [J]. Statistical Research, 2007(2): 81-83.

[6] Gaofeng Yue, Wang Liping, Liu Li, et al. Study on forecasting and forecasting of peony flowering period based on meteorological conditions [J]. Jiangxi Agricultural Journal, 2020, 32(2): 105—109.

[7] Li Ji, Xiaoran Liu, Qiang Wu, Qiang Li. Based on PCA to establish a prediction model for the early flowering of wintersweet [J]. Journal of Southwest Normal University (Natural Science Edition), 2022.10.009.

[8] Li Zhiqin, Du Jianqiang, Nie Bin, et al. A review of feature selection methods [J]. Computer Engineering and Applications, 2019, 55(24).

[9] Xuexue Yang, Qiang Liu. Economic loss prediction of storm surge disaster based on KPCA-RBF model [J]. Marine Science, 2021, 45(10): 32-39.

[10] Fu Hua, Fu Yu, Zhao Juncheng, Xu Tong, Lu Wanjie. Gas Emission Prediction Based on KPCA-ARIMA Algorithm [J]. Journal of Liaoning Technical University (Natural Science Edition), 2022, 41(05): 406 - 412.

[11] Gen Li, Guohua Zou, Xinyu Zhang. A review of high-dimensional model selection methods [J]. Mathematical Statistics and Management, 2012.04.004.

[12] H.Wang, X.Zhang and G.Zou (2009) Frequentist Model Averaging Estimation: A Review. Journal of Systems Science and Complexity, 22, 732-748.

[13] H.Liang, Zou., Wan, A.T.K. and Zhang, X. (2011) Optimal Weight Choice for Frequentist Model Average Estimators. Journal of the American Statistical Association, 106, 1053-1066.

[14] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996(24): 123-140.

[15] rong Zhu, Guohua Zou, Xinyu Zhang. Model Averaging Method for Partial Function Linear Models [J]. Systems Science and Mathematics, 2018, 38(07): 777-800.