Characteristic Selection and Prediction of Octane Number Loss in Gasoline Refinement Process

Wei Li[1,2,†], Jiali Yang[3,†], Peihao Yang[3]and Sheng Li[3,*]

[1] Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang), Zhanjiang 524088, China
[2] CNOOC China Limited ZhanJiang Branch, Zhanjiang, Guangdong, 524057, China
[3] School of Mathematics and Computer, Guangdong Ocean University, Zhanjiang Guangdong 524088, China
[†] These authors contributed equally to this work.

Abstract. In the refining process of gasoline, accurate prediction of the octane number loss is conducive to production management to ensure the octane content in gasoline. Therefore, the relevant research has important theoretical significance and application value. Aiming at the characteristics of octane number loss with few samples, high dimensions and non-linear of the octane number loss, this paper uses maximum information coefficient, recursive characteristic elimination and random forest regression algorithm to select the main characteristics, and establishes the octane number loss prediction model based on least squares support vector machine respectively. Compared with the three algorithms of support vector machine, BP neural network and ridge regression, the experimental results show that the two models of ridge regression and least square support vector machine have higher prediction accuracy, but the least square support vector machine has the best effect.

1 Introduction

Octane number (RON) is a digital expression to measure the anti-knock combustion ability of gasoline in automobile cylinder. It is one of the most important indexes to measure the quality of gasoline. The higher the value is, the better the its anti-knock performance is. An accurate prediction of RON loss can help production managers to understand production status and adjust production plan in time. RON loss data has the characteristics of few samples, high dimension and non-linearity. Generally speaking, not all the characteristics of such data play a decisive role in the target characteristics. Therefore, it is inefficient and inaccurate to predict the target characteristics directly [1],and it is necessary to select the main characteristics.

In order to improve the prediction efficiency and accuracy of target characteristics, the NOx emission prediction model based on deep cycle neural network is given in reference [2]. This model plays very well in both fitting effect and prediction accuracy, and can be effectively applied to the prediction of NOx concentration at the outlet of flue gas denitration system. A reaction kinetics model of MIP process based on structure oriented aggregation is proposed to predict and optimize the effects of outlet temperature and mass ratio of catalyst to oil on the reaction process [3]. In reference [4], support vector machine is used to predict obstructive sleep apnea in large-scale clinical samples in China, which provides a simple and accurate method for early identification of OSA patients.

In this paper, we use the maximum information coefficient, recursive characteristic elimination, and random forest regression [5] algorithm to select the main characteristics that affect RON loss, and use the least squares support vector machine model to predict and evaluate RON loss.

2 Related work

The main work of this paper includes: selecting the main characteristics of RON loss; establishing the prediction model of RON loss based on least squares support vector machine; testing the model. Figure 1 shows the process framework of characteristic selection and prediction of RON loss.

*Corresponding author's e-mail: lish_ls@gdou.edu.cn and lish_ls@sina.com

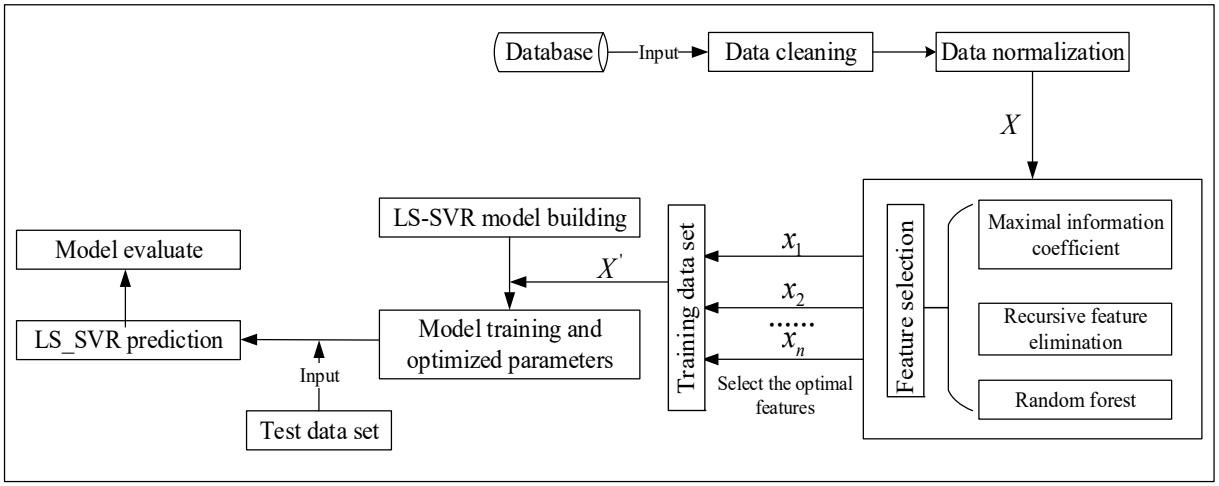

Fig.1 Process framework for octane number loss prediction

2.1 Characteristic selection

The so-called characteristic selection is to select meaningful characteristics, establish relevant data sets and input them into machine learning model for training. It is usually considered from two aspects. One is whether the characteristic is divergent, and the other is whether the characteristic is related to the target variable. The characteristic with high correlation should be given priority as the characteristic of the target variable [6].

According to the form of characteristic selection, characteristic selection methods can be divided into three kinds: filter, which scores each characteristic according to divergence or correlation, sets the threshold or the number of threshold to be selected to select characteristics; Wrapper, according to the objective function (usually the prediction effect score), selects several characteristics each time, or excludes some characteristics; Embedded method, which uses some machine learning algorithms and models for training to get the weight coefficient of each characteristic, and select characteristics according to the coefficient from large to small [7].

Characteristic selection is the process of selecting the best characteristic for the target variable from the original data. That is, for given characteristics, we select a subset of the best characteristics to improve the performance of machine learning in predicting the target variable. In this paper, three methods, maximum information coefficient (MIC), recursive characteristic elimination (RFE) and random forest (RF), are used to select the main characteristics that affect RON loss.

2.1.1 Maximum information coefficient (MIC)

The maximum information coefficient (MIC) is one of the filtering methods. It is used to measure the degree of correlation between two variables and, linear or nonlinear strength, and is often used in feature selection of machine learning. According to the nature of MIC, it has universality, fairness and symmetry to assist Pearson correlation coefficient in feature selection [8]. The concept of mutual information should be used to calculate the maximum information coefficient. The calculation formula of mutual information concept is shown in (1).

$$I(X;Y)=\sum_{j=1}^{m}\sum_{i=1}^{n}p(x_i,y_j)\log_2\frac{p(x_i,y_j)}{p(x_i)p(y_j)}, \qquad (1)$$

where $p(x_i,y_j)$ is the joint probability between variables X and Y. The calculation of joint probability is quite troublesome. To solve this problem, the calculation formula of MIC is given, such as (2).

$$MIC=\max_{|X||Y|<B}\frac{I[X;Y]}{\log_2(\min(|X|,|Y|))}, \qquad (2)$$

where B is a variable to be selected. It is selected in this paper to be is the 0.6 power of the amount of data.

2.1.2 Recursive feature elimination (RFE)

Recursive feature elimination (RFE) belongs to the packaging method. Feature recursion does not only look at the direct association between X and Y, but also judges whether the feature is suitable from the final performance of the model after adding the feature. Figure 2 is the flow chart of recursive feature elimination backward search algorithm. Firstly, all features are input into the model for classification, and the best features are selected. Then, the above selection work is repeated among the remaining features until the last feature classification is completed, and the best feature set is selected [9].

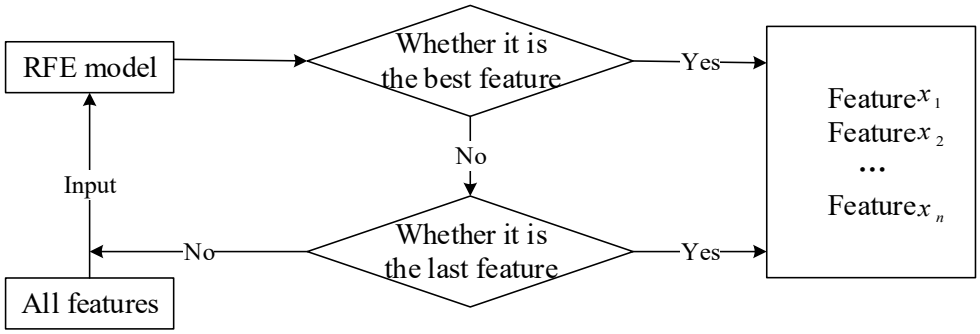

Fig.2 Recursive feature elimination algorithm flow

2.1.3 Random forest (RF)

The basic idea of random forest (RF) is to synthesize the regression results of multiple classification and regression tree (CART) to get the final result [10]. Firstly, random forest algorithm generates *n* sub datasets by self-help resampling, then randomly selects *K* features from these sub datasets, establishes CART, and repeats the above steps *m* times. Finally, the results of all decision trees are summarized and the weight is calculated, and the one with the highest weight is selected as the final result.

2.2 Establishment of least squares support vector machine (LS-SVR) model

Support vector machine, also known as support vector network, is a new machine learning method developed on the basis of statistical learning theory, which has the advantages of complete theory, strong adaptability, global optimization, short training time and good generalization performance [11]. Least squares support vector machine (*LS-SVR*) is a kind of support vector machine, which is obtained by transforming inequality constraints in standard support vector machine algorithm into equality constraints.

In *LS-SVR*, the optimization problem of support vector machine is replaced by equality constraint. Error variable is introduced to each sample e_i , and the $L2$ regularization term of error variable is added to the original function. The optimization formula of least squares support vector regression is shown in (3).

$$\min_{W,b} \frac{1}{2}\|W\|^2 + \frac{\lambda}{2}\sum_{i=1}^{m} e_i^2, \quad s.t. y_i(W.x_i + b) = 1, i = 1,2,...,m. \tag{3}$$

The square term of loss function is used to replace the insensitive loss function of support vector machine, and the inequality constraint with relaxed variables is replaced by the equality constraint with error variables. Usually, the dual problem can be obtained by using Lagrange multiplier method, and the Lagrange multiplier formula is introduced, such as (4).

$$L(w,b,e,a) = \frac{1}{2}\|W\|^2 + \frac{\lambda}{2}\sum_{i=1}^{m} e_i^2 + \sum_{i=1}^{m} a_i \{[y_i(W.\varphi(x_i)+b)] - 1 + e_i\}. \tag{4}$$

In this paper, the least square loss function is selected as the optimization index, then the function of *LS-SVR* model is (5):

$$f(x) = W.\varphi(x_i) + b = \sum_{i=1}^{m} a_i k(x_i, x) + b, \tag{5}$$

where $k(x_i, x)$ is a $m \times m$ kernel matrix, and we use the linear kernel function, as the following

$$k(x_i, x) = x_i^T x_j. \tag{6}$$

2.3 model evaluation index

After training and testing the model, we need to evaluate the model in order to judge the quality of the model. The model is evaluated from five aspects: mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE) and R2 coefficient of determination.

3 Experimental verification

3.1 software platform

The experimental operating platform is Ubuntu 64 bit operating system, python3.7, the compiler IDE is Spyder, the drawing platform is windows 10 64 bit operating system, and the drawing software is Visio 2013 and Adobe Illustrator (AI).

3.2 data sources and data processing

The data of this paper comes from Annex 1 and Annex 3 of question B in 2020 China graduate mathematical modeling competition. Data samples are collected from FCC gasoline refining unit. Annex 1 contains 325 data samples. Each sample has 354 operating variables and 14 characteristics of raw material property, product property, spent adsorbent and regenerated adsorbent, with a total of 368 characteristics. Annex 3 contains 285 and 313 data, with only 40 sample sizes and 354 operating variables.

First of all, we need to preprocess the data of 285 and 313 in Annex 3, and then add the processed data of 354 operation variables to the corresponding sample number in Annex 1. Then we need to normalize the data of 354 operation variables of 325 sample data in Annex 1. Because there is a large difference between different characteristics, the

calculated relationship coefficient will have a large difference, which can not be used in the optimization process Quickly find the optimal solution of the model. In this paper, we use the maximum minimum normalization method to map the features that affect RON loss to $[0,1]$, by the formula (7).

$$x_{ij}^{'} = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}, \ i = 1,2,...,n, \ j = 1,2,...,m. \quad (7)$$

where $x_{\min}$ is the minimum value of x_{ij}, $x_{\max}$ is the maximum value of x_{ij}, m is the data sample size, n is the feature dimension of the data.

3.3 Experimental results of feature selection

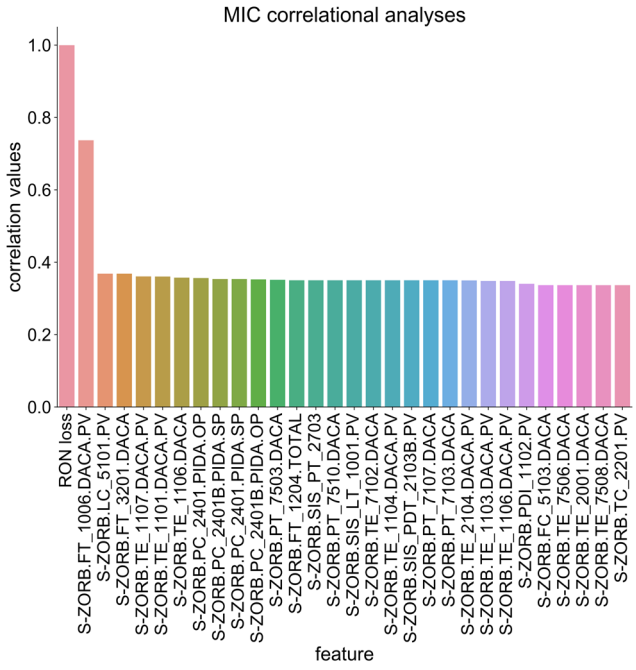

Fig.3 Top 29 features of maximum information coefficient

By using the maximum information coefficient, recursive feature elimination and random forest regression algorithm, the most important features and ranking information of each class are selected.

The maximum information coefficient is shown in the figure 3 above, and the top 29 features are selected for each characteristic.

The sum of the top 29 features of random forest weight is 0.38, as shown in the figure below:

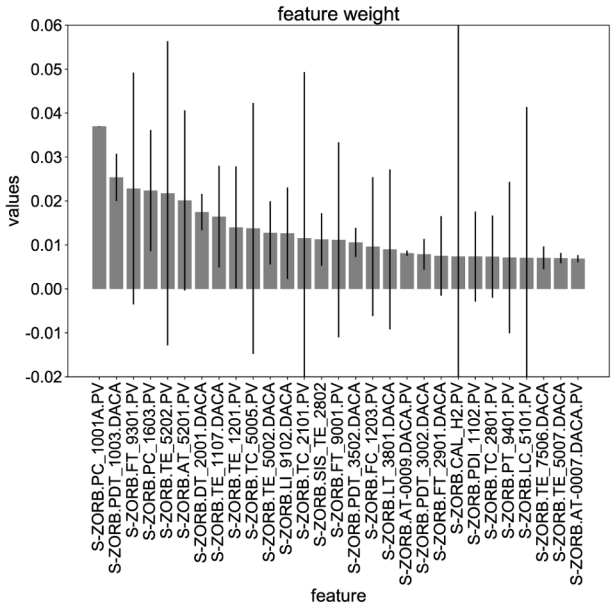

Fig.4 Top 29 features of random forest weights

The top 29 features extracted by the three algorithms mentioned above are shown in Table 1.

Table 1 Count the top 29 features of the three algorithms

Maximal Information Coefficient	Correlation	Random Forest	Weights	Recursive Feature Elimination	Rank
S-ZORB.FT_1006.DACA.PV	0.7370	S-ZORB.PC_1001A.PV	0.0370	S-ZORB.TE_6008.DACA.PV	1
S-ZORB.LC_5101.PV	0.3684	S-ZORB.PDT_1003.DACA	0.0254	S-ZORB.TE_6001.DACA.PV	2
S-ZORB.FT_3201.DACA	0.3684	S-ZORB.FT_9301.PV	0.0228	S-ZORB.PT_7508.DACA	3
S-ZORB.TE_1107.DACA.PV	0.3608	S-ZORB.PC_1603.PV	0.0224	S-ZORB.FT_1202.TOTAL	4
S-ZORB.TE_1101.DACA.PV	0.3605	S-ZORB.TE_5202.PV	0.0217	S-ZORB.SIS_PT_2703	5
S-ZORB.TE_1106.DACA	0.3574	S-ZORB.AT_5201.PV	0.0201	S-ZORB.TC_2201.PV	6
S-ZORB.PC_2401.PIDA.OP	0.3563	S-ZORB.DT_2001.DACA	0.0175	S-ZORB.PT_7103.DACA	7
S-ZORB.PC_2401B.PIDA.SP	0.3536	S-ZORB.TE_1107.DACA	0.0164	S-ZORB.PT_7505.DACA	8
S-ZORB.PC_2401.PIDA.SP	0.3536	S-ZORB.TE_1201.PV	0.0140	S-ZORB.PT_7510.DACA	9
S-ZORB.PC_2401B.PIDA.OP	0.3525	S-ZORB.TC_5005.PV	0.0138	S-ZORB.PT_7107.DACA	10
S-ZORB.PT_7503.DACA	0.3514	S-ZORB.TE_5002.DACA	0.0128	S-ZORB.SIS_LT_1001.PV	11
S-ZORB.FT_1204.TOTAL	0.3501	S-ZORB.LI_9102.DACA	0.0127	S-ZORB.FT_3302.DACA	12
S-ZORB.SIS_PT_2703	0.3501	S-ZORB.TC_2101.PV	0.0115	S-ZORB.LC_5102.DACA	13
S-ZORB.PT_7510.DACA	0.3501	S-ZORB.SIS_TE_2802	0.0113	S-ZORB.LC_5102.PIDA.PV	14

S-ZORB.SIS_LT_1001.PV	0.3501	S-ZORB.FT_9001.PV	0.0112	S-ZORB.PC_1001A.PV	15
S-ZORB.TE_7102.DACA	0.3501	S-ZORB.PDT_3502.DACA	0.0106	S-ZORB.CAL.CANGLIANG.PV	16
S-ZORB.TE_1104.DACA.PV	0.3501	S-ZORB.FC_1203.PV	0.0096	S-ZORB.PDT_2104.PV	17
S-ZORB.SIS_PDT_2103B.PV	0.3501	S-ZORB.LT_3801.DACA	0.0090	S-ZORB.TE_7102.DACA	18
S-ZORB.PT_7107.DACA	0.3501	S-ZORB.AT-0009.DACA.PV	0.0081	S-ZORB.FT_3301.TOTAL	19
S-ZORB.PT_7103.DACA	0.3501	S-ZORB.PDT_3002.DACA	0.0079	S-ZORB.FT_9403.TOTAL	20
S-ZORB.TE_2104.DACA.PV	0.3498	S-ZORB.FT_2901.DACA	0.0075	S-ZORB.FT_9001.TOTAL	21
S-ZORB.TE_1103.DACA.PV	0.3482	S-ZORB.CAL_H2.PV	0.0074	S-ZORB.FC_1101.TOTAL	22
S-ZORB.TE_1106.DACA.PV	0.3482	S-ZORB.PDI_1102.PV	0.0074	S-ZORB.FT_9402.TOTAL	23
S-ZORB.PDI_1102.PV	0.3403	S-ZORB.TC_2801.PV	0.0073	S-ZORB.FT_9302.TOTAL	24
S-ZORB.FC_5103.DACA	0.3366	S-ZORB.PT_9401.PV	0.0071	S-ZORB.TE_1104.DACA.PV	25
S-ZORB.TE_7506.DACA	0.3366	S-ZORB.LC_5101.PV	0.0071	S-ZORB.FT_9202.TOTAL	26
S-ZORB.TE_2001.DACA	0.3366	S-ZORB.TE_7506.DACA	0.0071	S-ZORB.FT_1504.TOTALIZERA.PV	27
S-ZORB.TE_7508.DACA	0.3366	S-ZORB.TE_5007.DACA	0.0070	S-ZORB.FT_1503.TOTALIZERA.PV	28
S-ZORB.TC_2201.PV	0.3366	S-ZORB.AT-0007.DACA.PV	0.0069	S-ZORB.FT_1001.TOTAL	29

Through the above three algorithms, the top five features of each method and a total of 4 common features are selected. A total of 19 features are selected. It is known from references [12,13] that sulfur content and alkane content also have certain influence on octane number loss. Eight characteristics including RON loss are selected from 14 characteristics of raw material property, product property, spent adsorbent and regenerated adsorbent, so 27 characteristics are finally selected, as shown in Table 2.

Table 2 Feature selection results

Feature	RF	MIC	RFE
S-ZORB.PC_1001A.PV	0.0370	null	15
S-ZORB.PDT_1003.DACA	0.0254	null	null
S-ZORB.FT_9301.PV	0.0228	null	null
S-ZORB.PC_1603.PV	0.0224	null	null
S-ZORB.TE_5202.PV	0.0217	null	null
S-ZORB.AT_5201.PV	0.0201	null	null
S-ZORB.DT_2001.DACA	0.0175	null	null
S-ZORB.FT_1006.DACA.PV	null	0.7370	null
S-ZORB.LC_5101.PV	null	0.3684	null
S-ZORB.FT_3201.DACA	null	0.3684	null
S-ZORB.TE_1107.DACA.PV	null	0.3608	null
S-ZORB.TE_1101.DACA.PV	null	0.3605	null
S-ZORB.SIS_PT_2703	null	0.3501	5
S-ZORB.PT_7510.DACA	null	0.3501	9
S-ZORB.SIS_LT_1001.PV	null	0.3501	11
S-ZORB.PT_7508.DACA	null	null	3
S-ZORB.FT_1202.TOTAL	null	null	4
S-ZORB.TE_6008.DACA.PV	null	null	1
S-ZORB.TE_6001.DACA.PV	null	null	2
Sulfur content of raw material, μg/g	null	null	null
Octane number of raw material	null	null	null
Saturated hydrocarbon (alkanes + cycloalkanes) ,v%	null	null	null
Olefin,v%	null	null	null
Aromatic hydrocarbon,v%	null	null	null
Sulfur content of product,μg/g	null	null	null
Product octane number	null	null	null
RON loss	null	null	null

3.4 Experimental results of octane number loss prediction

Feature selection is to provide data support for subsequent RON loss prediction, standardize the selected features, and then divide the 325 samples in Annex 1 into data, the first 265 samples as training set, and the last 60 samples as test set.

The BP neural network model is used for 15000 iterations, and the batch size is 16, as shown in Figure 5, where epochs represents the training times and loss represents the loss value. It can be seen that the loss value of model training and verification is infinitely close to 0, and tends to be stable.

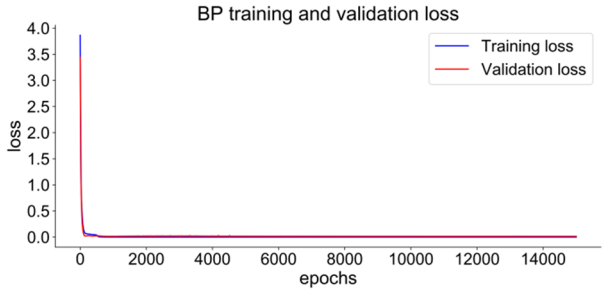

Fig.5 bp neural network training loss and verification loss

Figure 6 shows the prediction results with BP neural network. It can be clearly seen that the actual values and predicted values of the first 265 training samples are completely matched, and the actual values and predicted values of the last 60 test samples are not well matched.

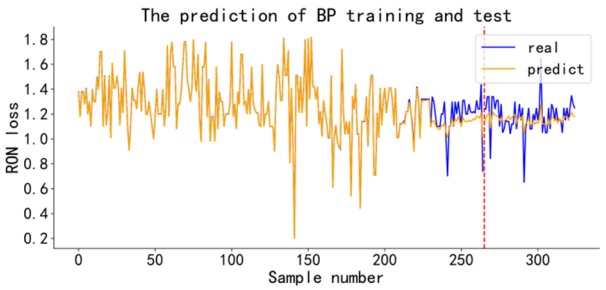

Fig.6 The prediction of octane number loss by BP neural network

Figure 7 shows the prediction of octane number loss by using support vector machine. It can be clearly seen that the actual value and predicted value of the first 265 training samples basically fit, and the actual value and predicted value of the last 60 test samples do not fit well.

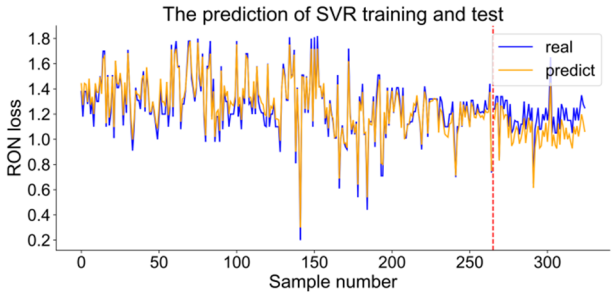

Fig.7 The prediction of octane number loss by SVR

Figure 8 and Figure 9 show the RON loss prediction results by ridge regression and least squares support vector machine algorithm respectively. It can be seen that the actual value and predicted value of the first 265 training samples are almost completely fitted, and the actual value and predicted value of the last 60 test samples are well fitted.

3.5 Evaluation index experimental results

In order to verify the quality of the model, it is verified from two aspects. Firstly, the least squares support vector machine uses linear kernel function and radial basis function kernel function to compare the mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE) and R2 determination coefficient; and four different algorithms are evaluated.

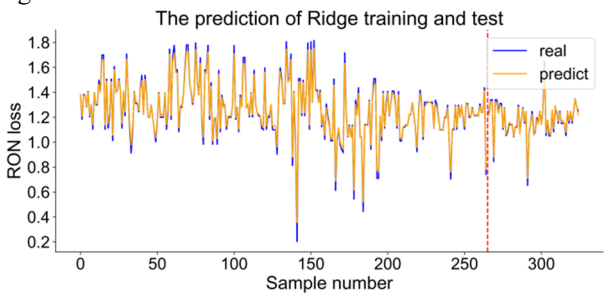

Fig.8 The prediction of octane number loss by Ridge

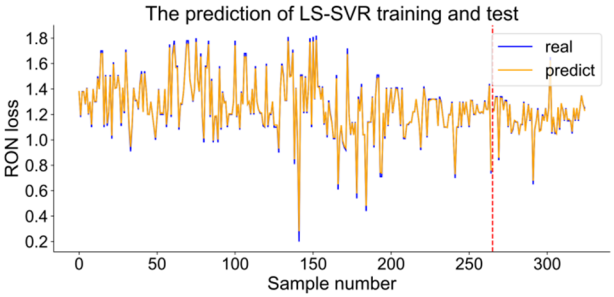

Fig.9 The prediction of octane number loss by least square support vector machine

Table 3 shows that linear kernel function is better than radial basis function in predicting RON loss.

Table 3 Evaluation index of least square support vector machine(test set)

Kernel function	MAE	MAPE	MSE	RMSE	R2
RBF	0.0958	8.8116	0.0156	0.1249	0.1426
Linear	0.0082	0.7158	0.0001	0.0104	0.9940

Table 4 and table 5 show that the least squares support vector machine is better than the other three models in predicting RON loss.

Table 4 Evaluation index for prediction of octane loss (training set)

Models	MAE	MAPE	MSE	RMSE
SVR	0.0428	3.7596	0.0026	0.0513
BPNN	0.0146	8.8116	0.0019	0.0443

Ridge	0.0208	2.0055	0.0008	0.0286
LS-SVR	0.0111	0.7158	0.0002	0.0153

Table 5 Evaluation index for prediction of octane loss (test set)

Models	MAE	MAPE	MSE	RMSE
SVR	0.1138	9.5944	0.0143	0.1198
BPNN	0.0811	9.1626	0.0811	0.1078
Ridge	0.0148	1.3258	0.0003	0.0184
LS-SVR	0.0082	0.7158	0.0001	0.0104

Figure 10 shows that the *R*2 values of training set and test set in ridge regression and least squares support vector machine are larger, indicating that these two algorithms are better, but the *R*2 score of least squares support vector machine is 1% higher than ridge regression.

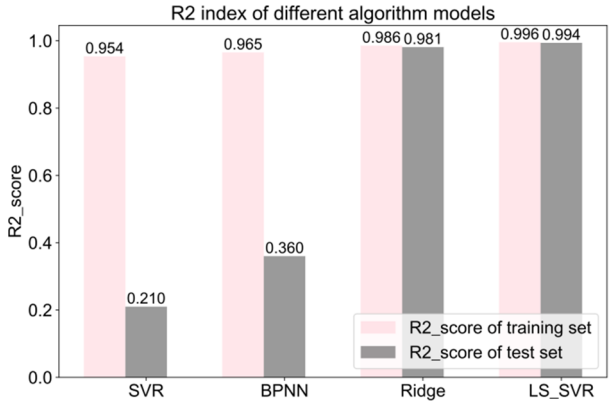

Fig.10 R2 index of four algorithm models

Through the statistics of the above model and kernel function evaluation index, we know that the least squares support vector machine using linear kernel function to predict RON loss is better than radial basis function kernel function, and the most prominent is that the R2 of RBF is only 14%; and the accuracy of *LS-SVR* model to predict RON loss is higher than other algorithms, and its *R*2 reaches 99%.

4 Conclusions

(1) The main characteristics of octane number loss were selected by maximum information coefficient, feature elimination and random forest regression algorithm.

(2) Octane number loss is predicted based on least squares support vector machine.

(3) Through the comparison of the evaluation index *LS-SVM* with support vector, BP neural network and ridge regression model, the results show that *LS-SVM* has better experimental effect.

Acknowledgments

This work was supported by the Fund of Southern Marine Science and Engineering Guangdong Laboratory (Zhanjiang) (ZJW-2019-04) and the Natural Science Foundation of Guangdong Province (2018A030307062).

References

1. Pan Y.Q.,Yang Q.L.Qang Y.W. et al. (2010) Feature extraction and classification prediction of microarray data. Proceedings of 2010 First International Conference on Cellular,Molecular Biology, Biophysics and Bioengineering(Volume 6).
2. .Qian H.,Chai T.T., Zhang C.F. (2020)NOx Emission Prediction Model of SC R Flue Gas Denitrification System based on Deep Recurrent Neural Network. J. Eng. Ther. Ener. Pow., 35(08): 77-84.
3. Liu J.C., Chen H., Pi Z.P., et al.(2017)Reaction kinetic model for catalytic cracking MIP technology using structure oriented lumping method II . Simulation of a commercial unit., Pet. Tec., 46(05):519-523.
4. Huang W. C.,Lee P. L., Liu Y. T., et al. (2020)Support vector machine prediction of obstructive sleep apnea in a large-scale Chinese clinical sample Sleep,43(7):7
5. Wang W.(2017)Performance of feature selection in several decision tree ensemble methods. Doctoral dissertation,Anhui:Anhui Normal University.
6. Zhou Z.H. Machine learning. (2016)Beijing: Tsinghua University Press, pp.121-139,298-300.
7. Li Z.Q., Du J.Q.,Nie B., et al.(2019)Comp. Eng. Appl.,55(24): 10-19.
8. Shao F.B.,Yang S.,G.,Sun B.C.,et al. (2019)The Big Data Analysis of Rail Equipment Accidents Based on the Maximal Information Coefficien. J. Transp. Sa. Secur.,12(7):959-976.
9. Chen Q.,Meng Z,P., Su R.(2020) WERFE: A Gene Selection Algorithm Based on Recursive Feature Elimination and Ensemble Strategy. Front. Bioeng. Biotech.,8: 496.
10. Hoffmeister D.,Herbrecht M.,Kramm T.,et al.(2020) Evaluation Of Random Forest-Based Analysis For The Gypsum Distribution In The Atacama Desert. LAGIRS 2020: 2020 Latin American GRSS & ISPRS Remote Sensing Conference..
11. Gao X.Q., Wang Z.W., Lu C., et al. (2020)Distribution network fault prediction model based on least squares support vector machine. J. Wuhan Univ. Tech.,42(01): 23-29.
12. Xu H.Y.(2020)Factors influencing octane number of gasoline. Brand. Stand.,05:49-50+52.
13. Sun S.K., Wang Z.Y., Cui Z., et al.(2020) Effect of process conditions on catalytic diesel hydroconversion. Eng. Ther. Ener. Pow., 48(08): 806-810.