Quality prediction of petroleum products based on KPCA-RNN

Jinhui Sun 1,a and Xuebin Sun 2

1 Key Laboratory of Universal Wireless Communications Ministry of Education, Beijing University of Post and Telecommunications, 10th Xitu Road, Haidian Strict, Beijing, China
2 Key Laboratory of Universal Wireless Communications Ministry of Education, Beijing University of Post and Telecommunications, 10th Xitu Road, Haidian Strict, Beijing, China

a Corresponding author: author@e-mail.org

Abstract. With the application of the Internet of Things in oil production, a large number of equipment parameters and environment parameters in production process can be collected. The data can be applied to production equipment troubleshooting and product quality prediction. Prior gasoline quality testing has strict requirements for equipment and operation, but with a low efficiency. The paper applies RNN in a new situation and proposes a gasoline production quality prediction model fused with dimensionality reduction which is based on improved RNN, to achieve the real-time monitoring for gasoline production quality. Since the collected data is time-series data, the data features are complex, high-dimensional, and contain a large amount of noise and outliers. Therefore, the data needs to be cleaned, and feature engineering is used to achieve feature selection and dimensionality reduction. The model is trained and validated by an actual data set of gasoline production in an offshore petroleum field. The results show that the established prediction model fused with dimensionality reduction technique and based on improved RNN can predict the quality of gasoline products in real time and ensure a high accuracy.

1. Introduction
The process of petroleum from mining to the refining and production of various products requires a series of complex processes. The adjustment of the production process is currently based on experience and the changes in equipment parameters. Petroleum is a complex, multi-component mixture with a wide boiling range from ambient temperature to above 500 °C. According to the characteristic, the current refining of petroleum products is mainly achieved by petroleum distillation.

In addition, the quality testing process of petroleum products is complex, and has high requirements for the equipment and testing technique. The paper takes gasoline products as an example to predict the quality of gasoline products, and regards the distillation point of gasoline as the prediction target. If all batches of products are sampled, it will take a lot of manpower and material resources.

With the application and development of the Internet of Things technology in the petrochemical industry, it is possible to collect data on all equipment states and environmental conditions in the petroleum extraction and refining process through a large number of sensors. The paper proposes to predict the quality of petroleum products based on the process parameters and environmental...
parameters. The data processing helps to reduce the time and space consumption of the model calculation, and improves the robustness and generalization performance of the model. Data processing is also a part the paper puts emphasis on.

PCA can transform a large number of correlated feature vectors into principal components containing a certain number of uncorrelated feature vectors. The goal of PCA is to reduce the dimensions of the initial variables while preserving as much information as possible.

RNN is a neural network that models sequence data. RNN create complex and nonlinear models through multiple layers of hidden layers. The hidden layers were once the cascades of nonlinear processing units and were responsible for feature extraction and conversion.

2. Related works
LU proposed an improved PLS algorithm to create a model to predict the distillation point of heavy diesel, and the process variables are taken as the input to the model after dimensionality reduction processing [1]. Li predicted the distillation point of aviation kerosene with partial least squares method [2]; Xue Xu applied the dynamic partial least square method to the prediction of gasoline distillation point. Static partial least square method can only extract static features so it is not suitable for dynamic distillation processes. The dynamic built-in partial least square method is proposed to predict the dynamic distillation process [3]. Qi Li proposed a hybrid model based on PLS and LS-SVM method to predict aviation kerosene quality prediction. Through the LS-SVM algorithm, the model has a fast convergence and high precision for the real-time monitoring of aviation kerosene quality [8]. Shum used a feed-forward neural network to build a classification model as well as an error backhaul training network. Thus, the model can accurately identify network attacks [4]. Skaruz applied RNN to network attack detection, and the RNN achieves higher classification accuracy in classification [5]. Meng's regression model combining nuclear PCA and RNN is successfully applied to cyber threat analysis [6]. Yu applied RNN to the regression predictions of weather. The weather detection parameters are time series data, and the prediction of weather using RNN improves the artificial weather prediction accuracy by 4.8% [7].

3. Algorithm
3.1. KPCA
The working principle of PCA is to select the direction with the largest variance one by one, and the covariance matrix is processed as following:

\[ \sum x_i x_i^T \]  
(1)

Eigenvalue decomposition is performed, and the feature vector \( \omega_1 \) corresponding to the largest eigenvalue is taken, then

\[ \sum x_i x_i^T - \lambda_1 \omega_1 \omega_1^T \]  
(2)

The eigenvalue decomposition is performed, and the eigenvector \( \omega_2 \) corresponding to the largest eigenvalue is taken. The dimension of the low-order space after dimensionality reduction can be directly set, or the threshold of the eigenvalue can be set to dynamically determine the dimension number of the space after dimension reduction.

The PCA dimensionality reduction can remove constant values and reduce noise, making the task of machine learning more precise, while identifying the main features from the data, and reducing the time and space complexity of the model. The data in this paper contains lots of nonlinear feature, so the paper applied KPCA to reduce dimension. The KPCA dimensionality reduction can remove constant values and reduce noise, making the task of machine learning more precise, while identifying the main features from the data, and reducing the time and space complexity of the model. Captions
should be typed in 9-point Times. They should be centred above the tables and flush left beneath the figures.

3.2 RNN
The purpose of the RNN is to process sequence data. In the traditional neural network model, from the input layer, the hidden layer to the output layer, the layers are fully connected, and the nodes between each layer are disconnected. But the common neural network is powerless for many problems. For example, the current output depends not only on the current output, but also on the "memory" of the previous input. In RNN, current output of a sequence is also related to the previous output. The specific expression form is that the network memorizes the previous information and applies it to calculate the current output; so the nodes between the hidden layers are connected, and the input of the hidden layer includes not only the output of the input layer, but also the output of the hidden layer at the previous moment. In theory, RNN can process sequence data of any length. However, in practice, in order to reduce complexity, it is often assumed that the current state is only related to the previous states in a short length. For any sequence index number \( t \), the hidden state \( h(t) \) is obtained by \( x(t) \) and \( h(t-1) \):

\[
h(t) = \sigma(z(t)) \\
= \sigma(Ux(t) + Wh(t-1) + b)
\]  

(3)

where \( \sigma \) is the activation function of RNN and \( b \) is the bias of the linear relationship.

The expression of the output \( o(t) \) of in the model when the sequence index number \( t \) is shown as:

\[
o(t) = Vh(t) + c
\]  

(4)

At the end of the sequence index number \( t \), the prediction output is:

\[
y^*(t) = \sigma(o(t))
\]  

(5)

Different from the traditional machine learning model, the hidden layer units are completely equal to each other, and the hidden layer in the middle of the RNN is sequential from left to right.

The RNN shares the same weight and generalizes in several time steps. Important features of RNN include:

1. Each step has an output, and RNN contains a recurrent connection among the hidden units as shown in Figure 1;
2. Each step produces an output, neural network with recurrent connection only exists between the output of the current moment and the hidden unit of the next moment.;
3. A recurrent connection exists among hidden units as shown in Figure 2, but the recurrent network produced a single output after reading the entire sequence.

The RNN model was is established to predict the quality of gasoline products. The error gradient is used to calculate the model gradient, and the adaptive learning rate algorithm is used to improve the convergence of the modal, at the same time use stochastic gradient descent algorithm to train the model.

The paper use regularization to solve the over-fitting problem with fixed training set size and fixed model complexity. L2 regularization is the most commonly used, which is based on the original loss function plus a penalty.

The regularized loss function makes balance between the model effect and model complexity. The trade-off between the two is measured by the regularization parameter \( \lambda \), which is small, the structural loss is more inclined to consider the model effect. On the contrary, it is more inclined to punish complex models. The purpose of the structured loss function is to make the weights in the model smaller to avoid over-fitting.
The regularization method limits the learning ability of the model by adding a parameter norm penalty $\Omega(\theta)$ to the objective function $J$.

![Figure 1. Structure of RNN.](image)

The activation function introduces a nonlinear relationship between neurons, which enables the model to better fit nonlinear relationships. The Sigmoid function is continuous, smooth, strictly monotonous, and symmetric with $(0, 0.5)$, and also a very good threshold function.

The RNN is trained with back propagation over time algorithm (BPTT). The BPTT algorithm is a training algorithm for the loop layers.

![Figure 2. Units in RNN.](image)

The gradient descent algorithm is a commonly-used method for solving unconstrained optimization problems and is simple to implement. The gradient descent method is an iterative algorithm, of which each step needs to solve the gradient vector of the objective function.

The stochastic gradient descent and its variants are likely to be the most widely used optimization algorithms in general machine learning, especially in deep learning.

4. Model Establishment

The distillation point of gasoline products is not only affected by the crude oil fractionation process, but also by the chemical cracking process. Therefore, the input parameters of the model are equipment parameters, environmental parameters and material information of petroleum distillation and chemical catalytic processes. The collected data contains a lot of information, the dimension is too high, and the features need to be filtered. The output of the model is the distillation point of the gasoline.

4.1. Data processing

Data cleaning includes the processing for the missing values, outliers, feature normalization and reduce noise contained in the data set.

The method of handling missing values is divided into two cases. In the first case, the data loss ratio exceeds 10%, so it is considered that the data can’t be correctly analysed and the data characteristics are discarded. In the second, if the data loss ratio is less than 10%, the data is considered to be credible, and the missing value is filled with the average or median of the existing data. The method is simple but beneficial to reduce the loss of sample information.

Abnormal values is also called outliers. After statistical analysis for the data, the points with a larger deviation from the data set will be determined as outliers. However, the large amount of data in petroleum production will enlarge the workload of the outlier detection with visual methods. The paper assumes that the data follows a normal distribution, so the outlier detection method based on normal distribution is used to identify outliers, and the average of the samples replaced the outliers.
Data normalization divides each feature vector into a uniform interval. Different data features have different dimensions and units, and the situation will affect the results of data analysis. In order to eliminate the dimensional influence among indicators, data normalization is needed to resolve the comparability between data indicators. After the original data has been standardized, the indicators are in a uniform order and ready for comprehensive comparison. It is believed that the data distribution is approximately Gaussian, so in order to provide a more data foundation for data dimensionality reduction, zero-mean normalization are used to normalize the data.

\[ z = \frac{x_i - \mu}{\delta} \]  

where \( \mu \) and \( \delta \) correspond to the mean and standard deviation of the features, respectively, and the quantized features obey the normal distribution and are distributed between [-1,1].

4.2. Dimensionality reduction
Each sample of the data set contains 8029 features. After statistical analysis, it is found that there are constants, variables with minimal information entropy in the data, and there are correlations among the variables. The correlations among variables will affect the analysis of model. Dimensionality reduction can effectively reduce the time complexity and space complexity of model calculation, effectively remove noise from datasets, and enhance the robustness of the model.

4.3. Model parameter settings
Learning Rate, weight initialization, number of layers, number of units, and regularization normalization are hyper-parameters of the model.

The parameter setting of the neural network lacks practical theoretical support, and is mainly debugged according to experience and practice. In order to obtain better parameter settings, the optimal parameters are obtained by comparing the same model parameters. In order to solve the problem, the learning rate of the training model is set to the adaptive learning rate to control the convergence speed and stability of the model as the gradient changes, and to improve the possibility of avoiding the local minimum of the model. As shown in Figure 4, the initialization and training method is given for the model. The number of hidden layer neurons and the number of layers in the hidden layer will affect the degree of fitting of the model. The dataset of the paper has many data features, complex structure and a high nonlinear strength. The more the number of neurons is, the better fitting of the model is, and the higher the prediction accuracy is, but it takes higher time and space consumption of training model and leads to a risk of over-fitting.
The learning rate uses adaptive learning rate, and the initial value is set to a small value. If the training error continues to decrease, the learning rate will not change. When the training error declines more slowly, the learning rate will increase, which helps the model to accelerate convergence.

![Figure 4. Predict values and real values in different models.](image)

The gradient of the model is calculated by error back-transition, and the model is trained with the stochastic gradient descent method.

### 4.4. Error measurement

The root mean square error (RMSE) is the square root of the ratio of the square sum of the deviation between the observed value and the true value and the number of observations $m$, which is used to measure the deviation between the observed values and the true values.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y'_i - y_i)^2}$$ (7)

Mean absolute error (MAE) is the average of the absolute error, which can better reflect the actual situation of the prediction value error.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y'_i - y_i|$$ (8)

### 5. Experiment Results

In order to effectively measure the accuracy of the prediction results of petroleum distillation points, the root mean square error and the average absolute error are used to reflect the accuracy of the regression model prediction results. The data set is divided into two parts, a training set and a test set. The data set is processed with uniform data, and each data contains 982 features. The training set contains 500 pieces of data, and the test set contains 100 pieces of data. After the model is trained, it is verified and compared on the same test set.

The experiments results (test error) are shown in Table 1 and Figure. 5.

| Method | RMSE     | MAE       |
|--------|----------|-----------|
| RNN    | 0.173456211325 | 0.153265332 |
| GBR    | 0.254963621353  | 0.201532561  |
| Huber  | 0.213452352163  | 0.170321523  |
| SVR    | 0.250362102347  | 0.196322135  |
The experimental results show that the test error of Hubor Regression is RMSE=0.2135. The test error of the improved model based on RNN proposed in the paper is RMSE=0.1734. It can be seen that the model proposed in the paper effectively improves the accuracy of the prediction results. At the same time, the data set without processed is tested in the model, and the training error is RMSE=0.3625. If without KPCA, the test error is RMSE=0.2741. At the same time, it spent about 3 times more time than using KPCA. It also can be seen that the processed data and KPCA can not only reduce the computational cost of the model, but also improve the accuracy and stability of the model prediction.

6.Conclusions
For the processing and forecasting of the collected time series data, the paper has referred to the idea of partial least square method, and proposed an improved RNN model with KPCA to realize online prediction of gasoline product quality. The experimental results showed that the improved RNN prediction model proposed in the paper can quickly and accurately predict the gasoline distillation point, and the model has a strong robustness and generalization performance. The paper has trained the processed and unprocessed datasets separately. The processing method of the data can effectively remove noise and abnormal points, supplement the vacancy value, and improve the accuracy of model prediction. The paper also compared the model with KPCA and without KPCA, and has found that KPCA can not only improve the accuracy but also save training time. Comparing the proposed model with the robust regression, the superiority of the improved RNN prediction model has been proved. In summary, the pre-processing of data and KPCA are necessary to guarantee the accuracy and stability of the model. At the same time, RNN has superiority in the processing of time series data, and has a good development prospect in the prediction of industrial production quality. The hyper-parameter setting will still be the direction for future improvement.

Acknowledgments
Thanks to Professor Xuebin Sun for teaching and helping me. With the help of him, the establishment of the gasoline distillation point prediction model was achieved.
References

[1] W. X. Lu and D.X. Huang. Improved soft—sensor method and its applications to crude unit product quality. Control Engineering of China, vol. 11, on. 4, pp.296-300A(2008)

[2] Y. F. Li and F. Ren. Soft sensor of applied aviation kerosene dry point based on partial least squares regression of spline transformation. Computes and Applied Chemistry, vol. 26, no. 3, pp. 300-303(2009)

[3] Xue Xu, Qiang Liu and Jinliang Din. Gasoline Dry Point Prediction of Fractionation Processes Using Dynamic Inner Partial Least Squares. 2017 11th Asian Control Conference (ASCC) Gold Coast Convention Centre, Australia December 17-20(2017)

[4] J. Shum and H.A. Malki. Network intrusion detection system using neural network In Proc. IEEE Fourth Int. Conf. on Natural Computation , pp.242-246(2008)

[5] J. Skaruz and F. Seredynski. Recurrent neural networks towards detection of sql attacks. In Parallel and Distributed Processing Symposium, pp.18(2007)

[6] Fanzhi Meng, Yunsheng Fu, and Fang Lou. A network threat analysis method combined with kernel PCA and LSTM-RNN. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) March 29–31(2018)

[7] Chanjong Yu, Heewoong Ahn and Junhee Seok, Coordinate-RNN for Error Correction on Numerical Weather Prediction . 2018 International Conference on Electronics, Information, and Communication (ICEIC). DOI: 10.23919/ELINFOCOM.2018.8330699(2018)

[8] Qi Li, Mengqi Xie and Xiaodong Du. Soft Sensing for Dry Point of the Aviation Kerosene Based on PLS and LS-SVM Method. 2017 Chinese Automation Congress (CAC). DOI: 10.1109/CAC.2017.8243417(2016)