Classification and prediction of wear state based on unsupervised learning of online monitoring data of lubricating oil

YONGSHENG SU1,a, DAYANG LI2,b, HENG WANG2, CHEN HE2, KUN YANG2, e*

1College of Power Engineering, Naval University of Engineering, Wuhan, 430033, China
2School of Energy and Power Engineering, Wuhan University of Technology, Wuhan, 430000, China
*aemail: suyongsheng1981@163.com, bemail: dieyoungli@qq.com,
*eCorresponding author’s e-mail: kyang@whut.edu.cn

Abstract. With the development of the times, the demand for high efficiency and reliability of machine performance in China's industry has become higher than ever before, and the traditional equipment condition evaluation method has also encountered a huge challenge. As an important branch of machine learning, cluster analysis is widely used in fault diagnosis and other fields because of its advantages of no prior knowledge and massive data processing. This paper introduces the concept of principal component analysis and K-means method based on time series in the field of oil data analysis. Through the test data on the oil monitoring data of the steam turbine unit in the power plant, the condition evaluation and classification of the equipment are carried out, among which the classification effect of 6μm particles, dielectric constant and 4μm particles is very obvious, and the water content is relatively obvious, which can basically distinguish each state. At the same time, on the basis of feature extraction, use the special cyclic neural network (RNN) in deep learning-long and short-term memory network (LSTM) to build a model of data for data prediction, under the powerful ability of LSTM to process temporal data, they have unearthed more irregular and non-linear trends in a large amount of historical data, and achieved better prediction results than traditional methods (ARIMA). Through clustering and prediction of lubrication parameter data, it can realize early warning and abnormal diagnosis of lubrication status and reduce damage to machinery and equipment.

1. Introduction

With the rapid improvement of comprehensive national strength in the past 40 years, China's traditional industries are developing vigorously in the direction of integration and intelligence. Correspondingly, the demand for mechanical equipment is increasing, and the design structure of mechanical equipment is becoming more complex, large-scale, diversified and precise. Therefore, the technical difficulty of condition monitoring and reliability evaluation of these equipment is becoming more and more high. The fault diagnosis and condition monitoring of mechanical equipment has become a new hot topic in various research fields.

As a branch of artificial neural network in machine learning, recurrent neural network (RNN) can combine previous learning information with current input to generate current output. Therefore, the
output of RNN is not only determined by the current input, but also related to the previous input, which makes RNN one of the powerful tools for non-stationary time series prediction[1]. It allows the design of such a model, the model is composed of multiple processing layers, which have the ability to model temporal data prediction[2]. However, the traditional RNN shows practical difficulties when the training network suddenly faces a long input/output sequence. Therefore, a gradient based method called long-term short-term memory neural network (LSTM) is used to develop a stable cyclic neural network structure. This new technology replaces the traditional RNN for time series prediction. The technology solves the problem of gradient disappearance/explosion. The units in the LSTM deep network structure can provide a powerful model for processing temporal data and forget the past information. In the case of ignoring the current input, it provides more flexibility for machine learning algorithm[3].

2. Principal Component Analysis

Principal component analysis (PCA) was first proposed by British mathematician Karl Pearson in the early 20th century[4]. It is one of the most commonly used dimensionality reduction methods. It is a linear method, which means that the transformation between the original data and the new low dimensional representation is a linear projection. Its main purpose is to reduce dimension, but it can also be used to explore the relationship between variables. Usually, before using another statistical method (such as regression or clustering), it is used as a data preprocessing method for data orthogonalization and eliminating redundancy caused by variable correlation.

The essence of principal component analysis is to extract orthogonal hidden variables from multidimensional data and maximize the variance of extracted variables. Take Fig.1 as an example. The two graphs show exactly the same data, but fig.1(b) reflects the data transformed by PCA. At this time, the new coordinate axis is the principal component we need.

![Fig.1. (a) Original Data (b) PCA Transformed Data](image)

If the x-axis in Fig.1 (b) is selected as the principal component, then the y-axis is the residual. Using principal component analysis, the measurement space can be decomposed into the principal subspace with the largest difference and the residual subspace. Given the m-dimensional parameter variable $X = (x_1 + x_1 + \cdots + x_m)(X \in \mathbb{R}^{m \times n})$, it needs to be transformed into another matrix $T$.

$$
\begin{align*}
    t_1 &= p_{11}x_1 + p_{21}x_2 + \cdots + p_{m1}x_m \\
    t_2 &= p_{12}x_1 + p_{22}x_2 + \cdots + p_{m2}x_m \\
    \vdots \\
    t_n &= p_{1n}x_1 + p_{2n}x_2 + \cdots + p_{mn}x_m
\end{align*}
$$

The form of matrix is as follows:

$$
\begin{bmatrix}
    p_{11} & p_{21} & \cdots & p_{m1} \\
    p_{12} & p_{22} & \cdots & p_{m2} \\
    \vdots & \vdots & \ddots & \vdots \\
    p_{1n} & p_{2n} & \cdots & p_{mn}
\end{bmatrix}
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_m
\end{bmatrix}
$$

$p_i, x_j \in \mathbb{R}^m, p_i, x_j$ is the standard Euclidean inner product, that is, the original data $X$ is projected onto the column of $P$, and the row of $T$ is the new base of column $p = \{p_1, p_2, \cdots, p_m\}$, which is
orthogonal and will become the principal component direction. The variance of $T$ decreases strictly. If there is a large correlation between some dimensions, the variance of $T$ will also concentrate on the first few principal components, while the variance of the latter principal components will become very small, even can be ignored. Principal component analysis not only retains most of the original information, but also reduces the dimensions by retaining the first few principal components with large variance.

After solving all the eigenvectors, it is very important to determine the number of principal components reasonably. The main consideration is that the dimension of the original data can be effectively reduced and the original data information can be retained to the maximum extent. Qin thinks that the best number of principal components is needed to minimize the reconstruction error[5]. This idea only needs to analyse the data, and does not need any prior knowledge. However, it lacks pertinence when monitoring specific known processes. Wang Haiqing and others improved this idea[6]. If the fault set is known, the number of principal components that can accurately detect the critical fault value of the principal component space and the residual space is selected when selecting the principal component. The two methods have their own advantages and disadvantages, which should be selected according to the actual situation. Generally, the cumulative percentage of variance (CPV) method is used, which is defined as follows:

$$CPV = \frac{\sum_{i=1}^{n} \lambda_i}{\text{tr}(\Sigma)}$$

(3)

When CPV>85%, $n$ is the number of principal components. This method is simple, widely used and reliable, but it is subjective and needs special analysis in some application scenarios.

In this paper, the data obtained from oil monitoring data of steam turbine unit of power plant are used to record the data obtained from 1 minutes in 13 hours from 17 hours in April 17, 2014 to 18 days, and the number of principal elements is 4.

3. Unsupervised Learning Clustering

A typical example of unsupervised learning is cluster analysis. Clustering analysis is the process of dividing data points into different clusters according to different methods. As an important branch of machine learning, clustering analysis is widely used in fault diagnosis and other fields. Clustering is an unsupervised learning process, that is to say, before dividing the data samples, we cannot understand the internal structure of the data completely, or we cannot need to know the number of categories of the data. The clustering algorithm will divide the categories according to the similarity between the data objects, without prior knowledge, which is more convenient and faster to deal with the modern society "Massive data".

This paper uses a time series clustering algorithm based on shape-based distance (SBD). First, the coefficient normalization method is used, and then the distance is calculated with the following formula.

$$SBD \left( \vec{x}, \vec{y} \right) = 1 - \max_w \left( \frac{C_{lw}(\vec{x}, \vec{y})}{R_l(\vec{x}, \vec{x}) + R_l(\vec{y}, \vec{y})} \right)$$

(4)

According to the expert evaluation model, the lubricating oil parameters are divided into five categories (failure, serious failure, mild failure, sub-health, health), and finally Fig.2 is obtained.
4. LSTM Prediction
Long-short term memory (LSTM) is a kind of time cyclic neural network. It is specially designed to solve the long-term dependence problem of general RNN. All RNNs have a chain form of repetitive neural network modules.
To explain how LSTM works, the loop part of the network is expanded into a directed acyclic graph. In each time step \( t(1 \leq t \leq n) \), the loop unit of each layer in the network will process the input data. The output of a specific time is predicted by the input data \( x_t \), the input data \( x = [x_1, x_2, \ldots, x_n] \) is composed of a series of independent variables which are processed in sequence for the last \( n \) consecutive time steps. The output of the last cycle layer (the second layer in this example) and the last time step \( x_n \) are fed into a dense layer to calculate the final prediction \( y \).

**Fig. 3. A General Example of A Two-Layer Recurrent Neural Network Expanded Over Time**

**Fig. 4. (a) Internal Logic Diagram Of Traditional Rnn Unit; (b) Internal Logic Diagram Of Lstm Unit**

The difference between traditional RNN and LSTM lies in the internal processing logic of the loop unit shown in Fig. 4. In the traditional RNN unit, there is only one internal state \( h(t) \). In each time step, the following formula is used to recalculate:

\[
g(h(t)) = g(W x_t + U h_{t-1} + b)
\]  

\( g(\cdot) \) is the activation function, and the hyperbolic tangent function is usually selected. \( W \) and \( U \) are adjustable weight matrices of hidden state \( h \) and input \( x \) respectively, and \( b \) is adjustable deviation vector. In the first time step, the hidden state is initialized to a zero vector whose length is a user-defined network super parameter.

In contrast, the LSTM has an additional state unit, or cell memory \( c_t \), that can store information, which controls the gate of information flow within the LSTM cell. The first gate is the forgetting gate, which elements of the unit state vector \( c_{t-1} \) are to be forgotten (to what extent).

\[
f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)
\]  

\( f_t \) is a vector with a range of \((0,1)\), \( \sigma(\cdot) \) is a logarithmic logic function, \( W_f, U_f \) and \( B_f \) define a set of learnable parameters of forgetting gate, that is, two adjustable weight matrices and a deviation vector.
In the next step, the potential update vector of cell state is calculated according to the current input \( x_t \) and the last hidden state \( h_{t-1} \), which is given by the following formula:

\[
\tilde{c}_t = \tanh (W_c x_t + U_c h_{t-1} + b_c)
\]  

(7)

\( \tilde{c}_t \) is a vector with a range of \((-1,1)\), \( \tanh(\cdot) \) is a hyperbolic tangent function, and \( W_c, U_c \) and \( b_c \) define another set of learnable parameters.

The second gate is the input gate for calculation, which (and to what extent) \( \tilde{c}_t \) information is used to update the cell state in the current time step:

\[
i_t = \sigma (W_i x_t + U_i h_{t-1} + b_i)
\]  

(8)

\( f_t \) is a vector with a range of \((0,1)\), \( \sigma(\cdot) \) is a logarithmic logic function, \( W_i, U_i \) and \( b_i \) define a set of learnable parameters of forgetting gate, that is, two adjustable weight matrices and a deviation vector.

In the next step, the potential update vector of cell state is calculated according to the current input \( x_t \) and the last hidden state \( h_{t-1} \), which is given by the following formula:

\[
\tilde{c}_t = \tanh (W_c x_t + U_c h_{t-1} + b_c)
\]  

(9)

\( \tilde{c}_t \) is a vector with a range of \((-1,1)\), \( \tanh(\cdot) \) is a hyperbolic tangent function, and \( W_c, U_c \) and \( b_c \) define another set of learnable parameters.

The second gate is the input gate for calculation, which (and to what extent) \( \tilde{c}_t \) information is used to update the cell state in the current time step:

\[
i_t = \sigma (W_i x_t + U_i h_{t-1} + b_i)
\]  

(10)

\( i_t \) is a vector with a range of \((0,1)\), \( \sigma(\cdot) \) is a logarithmic logic function, and \( W_i, U_i \) and \( b_i \) define a set of learnable parameters of the input gate.

Combined with the above formula, the unit state update formula is as follows:

\[
c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t
\]  

(11)

\( \odot \) is a scalar multiplication because both the vector \( f_t \) and the range \( i_t \) are in \((0,1)\). Formula (3-8) can be explained by its definition: when the value of \( f_t \) is about 0, the information stored in \( c_{t-1} \) will be forgotten; when the value of \( f_t \) is about 1, the information stored in \( c_{t-1} \) will be retained. Similarly, \( i_t \) determines that new information stored in \( \tilde{c}_t \) is added to the cell state \( i_t \) value is about 1 and ignore information \( i_t \) value is about 0. Like the hidden state vector, the cell state is initialized by the zero vector in the first time step. Its length corresponds to the length of the hidden state vector.

The third and final gate is the output gate, which controls the information flowing into the unit state \( c_t \) of the new hidden state \( h(t) \). The calculation formula of output gate is as follows:

\[
o_t = \sigma (W_o x_t + U_o h_{t-1} + b_o)
\]  

(12)

\( o_t \) is a vector with a range of \((0,1)\), and \( W_o, U_o \) and \( b_o \) define a set of learnable parameters of the output gate. From this vector, the new hidden state \( h(t) \) can be calculated. The formula is as follows:

\[
h_t = \tanh \left( c_t \odot o_t \right)
\]  

(13)

Obviously, unit state \( c_t \) can efficiently learn long-term dependence. Because its linear interaction with other LSTM cells is very simple, it can store unchanged information for a long time. In the training process, this feature helps to prevent the gradient explosion or gradient disappearance in the backward propagation step. Like other neural networks, a layer can be composed of multiple units (or neurons), and the cell length and hidden state vector in LSTM can be freely selected. In addition, we can stack multiple layers together. The last LSTM layer output of the last time step \( h_n \) is connected to a single output neuron through the dense layer, which calculates the final prediction result. The calculation formula of dense layer is as follows:

\[
y = W_d h_n + b_d
\]  

(14)

\( W_d \) is the weight matrix of dense layer and \( b_d \) is the bias term.

In order to test the prediction accuracy of the model, a set of initial parameters is set for the model. The parameters of LSTM model are initialized as follows: 90% of the data is used as the training set, and 10% of the data is used as the test set. The training set containing 90% data is selected as the verification set, and the training step size is initially set to 10. For the super parameters of LSTM model: the number of hidden layer neural network units is set to 20, the number of iterations is 50, and the
learning rate is 0.01. After setting the parameters, the training of the model is completed, and the average root mean square error (RMSE) is obtained by using the 50 fold cross validation algorithm. The experimental results are shown in Fig. 5.

![Fig. 5](attachment:image.png)

**Fig. 5.** (a) The RMSE index of LSTM varies with the number of iterations; and (b) The prediction effect of LSTM

As can be seen in Fig. 5(a), the error curve of the training set always has a slope during the whole iteration process, which means that it has not yet converged after the iteration, while the error curve of the training set is rapidly parallel to the axis when the training set has no more than 10 iterations. When the convergence speed of the test set is much faster than that of the training set, the problem of overfitting may occur in the model. As can be seen from Fig. 5(b), the trend of the predicted value is basically consistent with the trend of the actual value, but the fluctuation range is smaller than the actual value, so a more ideal prediction model is basically obtained.

The results obtained by using the traditional Arima prediction method are shown in the figure. It can be seen that the prediction result of LSTM is better than the original curve.

![Fig. 6](attachment:image.png)

**Fig. 6.** Prediction results of Arima algorithm
5. Conclusion
In this paper, the principal component analysis of the data is carried out, and then the clustering algorithm based on shape based distance (SBD) is used for clustering. Finally, the LSTM algorithm is used to predict the data. Compared with the traditional ARIMA algorithm, the fitting effect of the original curve is better. By using unsupervised clustering algorithm, the subjective factors in setting threshold are reduced. At the same time, the accurate prediction ability of LSTM algorithm will provide a strong theoretical basis for the fault processing technology of oil online monitoring, and improve the reliability of oil online monitoring technology. By clustering and predicting the lubrication parameter data, early warning and abnormal diagnosis of lubrication state are realized, and the damage of machinery and equipment is reduced.

Acknowledgments
The authors would like to thank the National Natural Science Foundation of China (No. 52071241) for their support in this research.

References
[1] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training Recurrent Neural Networks[C]// Proceedings of the 30th International Conference on Machine Learning. 2013.
[2] Perra N, Gonçalves B. Modeling and predicting human infectious diseases[M]. Social Phenomena, 2015:59–83.
[3] Wang X, Liangyu M A, Wang B, et al. A hybrid optimization-based recurrent neural network for real-time data prediction[J]. Neurocomputing, 2013, 120(10): 547-559.
[4] Pearson K. Olines and planes of closest fit to systems of points in space[J]. Edinburgh and Dublin Philosophical Magazine and Journal of Science, 1901,6:572-575.
[5] Qin S Joe, Dunia Ricardo. Determining the number of principal components for best reconstruction [J]. Journal of Process Control 2000(10): 245-250.
[6] Haiqing Wang, Shiming Yu. Method for selecting the number of principal components based on fault diagnosis performance optimization [J]. Acta Chem Sinica, 2004, 55 (2): 214-219.