Correlation Coefficient Based Cluster Data Preprocessing and LSTM Prediction Model for Time Series Data in Large Aircraft Test Flights

Hanlin Zhu[1], Yongxin Zhu[1(✉)], Di Wu[1], Hui Wang[1(✉)], Li Tian[1], Wei Mao[2], Can Feng[2], Xiaowen Zha[2], Guobao Deng[2], Jiayi Chen[2], Tao Liu[2], Xinyu Niu[3], Kuen Hung Tsoi[3], and Wayne Luk[4]

[1] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China
{zhuyongxin, wanghui}@sari.ac.cn
[2] Commercial Aircraft Corporation of China Ltd., Shanghai, China
[3] Shenzhen Corerain Technologies Co. Ltd., Shenzhen, China
[4] Imperial College London, London, UK

Abstract. The Long Short-Term Memory (LSTM) model has been applied in recent years to handle time series data in multiple application domains, such as speech recognition and financial prediction. While the LSTM prediction model has shown promise in anomaly detection in previous research, uncorrelated features can lead to unsatisfactory analysis result and can complicate the prediction model due to the curse of dimensionality. This paper proposes a novel method of clustering and predicting multidimensional aircraft time series. The purpose is to detect anomalies in flight vibration in the form of high dimensional data series, which are collected by dozens of sensors during test flights of large aircraft. The new method is based on calculating the Spearman's rank correlation coefficient between two series, and on a hierarchical clustering method to cluster related time series. Monotonically similar series are gathered together and each cluster of series is trained to predict independently. Thus series which are uncorrelated or of low relevance do not influence each other in the LSTM prediction model. The experimental results on COMAC's (Commercial Aircraft Corporation of China Ltd) C919 flight test data show that our method of combining clustering and LSTM model significantly reduces the root mean square error of predicted results.

Keywords: Cluster · Time series · Correlation coefficient · LSTM

1 Introduction

As a recent development of Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) network has been applied to handle time series data in multiple domains such as speech recognition and financial prediction in recent years. LSTM often achieve high accuracy in many problems by containing a memory cell that can remember long term dependencies. A typical LSTM cell contains 4 gates, each with

M. Qiu (Ed.): SmartCom 2018, LNCS 11344, pp. 376–385, 2018.
https://doi.org/10.1007/978-3-030-05755-8_37

their own weights and biases, leading to a high computational cost during inference among time series data.

Real-time analysis of time series data is required in aircraft test flights as the safety and stability is of great concern in airplane. In anomaly detection for aircraft, acceleration data which are collected via dozens of sensors play an important role in real-time prediction and diagnosis. Each sensor can provide a high sampling frequency time sequence and we can get a high dimensional series through a large number of sensors. The purpose is to use past data to predict future data. If the gap between predicted data and measured data exceeds a threshold value, the position of sensor which records such data may trigger an anomaly at this time.

During the test flight of the COMAC (Commercial Aircraft Corporation of China, Limited) C919 airplane, terabytes of data are collected. The data have a character of high dimension and high frequency. In previous research, the long short term memory network has shown good performance in such kind of big data. However, the high dimensionality of the data complicates their architecture.

Moreover, uncorrelated features can lead to unsatisfactory analysis and complicated prediction models, due to the curse of dimensionality. Direct deployment of LSTM prediction model in previous research fails to ensure a satisfactory prediction performance.

To minimize the impact of uncorrelated features on time series data, we propose a novel method of clustering and predicting multidimensional aircraft time series to detect anomalies in flight vibration time series in this paper. Clustering is a branch of unsupervised machine learning. In clustering, different metric functions are used to measure the distance or similarity between data or clusters, so that close data or similar data can be gathered together.

In the area of time series analysis, traditional ARMA and GARCH model have limit in dealing with the high dimensional and complex problem. LSTM neural networks overcome the vanishing gradient problem through recurrent neural networks (RNNs) by employing multiplicative gates that enforce constant error flow through the internal states of special units called ‘memory cells’. Because of the ability to learn long term correlations in a sequence, LSTM networks obviate the need for a pre-specified time window and are capable of accurately modeling complex multivariate sequences.

The contributions of this paper can be summarized as follows.

(1) A prediction model which combines clustering and LSTM together to minimize the impacts of uncorrelated features in time series data.
(2) Hierarchical clustering based on Spearman’s rank correlation coefficient between two series to gather the monotonically similar series, and to remove the unrelated ones.
(3) Modification of the LSTM model for training and prediction in each cluster filtered in the clustering stage.
(4) Evaluation of our method with COMAC’s C919 flight test.

The following outlines the rest of this paper. Section 2 introduces related work. Section 3 presents the basic algorithm and our combination model. Section 4 shows the experimental results and discussion. Section 5 draws a brief conclusion.

2 Related Work

There are many studies focusing on series tendency analysis, time series clustering and time series prediction.

Cao et al. [1] introduced a framework of real-time anomaly detection for flight testing. They used this method to solve other kinds of similar problems based on transfer learning. Their approach is to establish an anomaly detection model for dangerous actions of aircraft testing fights.

Hsu et al. [2] introduced a feature selection method through Pearson's correlation coefficient clustering. They used Pearson's correlation coefficient to measure the similarity between variables and clustered the features through hierarchical clustering. They used the UCI Arrhythmia dataset and SVM algorithm for experiment and analysis the validity of the method.

Gauthier [3] introduced a trends detection method through Spearman's rank correlation coefficient. He used the Spearman's rank correlation coefficient and Mann-Kendall test to analysis the similarity of MTBE data.

In the relevant area, cloudlet-based mobile cloud computing [8], reinforcement learning [12], K-means and PCA algorithm [6] are also used to analyze the data. In order to improve the efficiency of status detection, the FPGAs are used to accelerate Genetic programming [7], one-class SVM [9] and some other algorithms.

3 LSTM Joint Cluster Architecture

3.1 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient or Spearman's rho is similar to Pearson's correlation, which can be used to measure how well the relationship between two variables. It is a nonparametric measure of rank correlation. The difference between Spearman correlation and Pearson correlation is that the former can assess monotonic relationships (whether linear or not).

Suppose we have two series x_i and y_i, The Spearman's rank correlation coefficient can be calculated through the following equation:

$$\mathrm{r}_s = 1 - \frac{6\sum d_i^2}{n(n-1)} \tag{1}$$

where x_i is the difference between ranks for each x_i, y_i data pair and is the number of data pairs.

Spearman's coefficient can be applied to both continuous and discrete data. When there are no repeated data and each variables is a perfect monotone function of the other, a perfect Spearman correlation occurs. If data have a similar rank, Spearman correlation will get close to +1. On the contrary, if data have a dissimilar rank, it will decline to −1. Besides, if two series aren't related, it will close to 0.

3.2 Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering establishes a nested clustering tree according to the degree of similarity between data objects that do not belong to the same category [5]. Firstly, Hierarchical Agglomerative Clustering uses each raw data point as one class, and calculate the distance between different classes of data points. The smaller the distance, the higher the similarity. In our work, the closer the absolute value of correlation coefficient is to 1, the closer the two sequences are. This techniques involve aggregating the categories with the smallest distance and iterating through the process, until the number of categories reaches the expected value or other termination conditions are met.

There are three ways to calculate the distance between two categories of data: Single Linkage, Complete Linkage and Average Linkage. Single Linkage takes the distance between two data objects with the smallest distance among the data objects belonging to different categories as the distance between the data objects of the two categories. Complete Linkage just does the opposite, which takes the largest distance as the distance between different categories.

3.3 LSTM Prediction Model in Time Series

The Long Short-Term Memory (LSTM) architecture, which uses purpose-built memory cells to store information, is good at finding and exploiting long range dependencies in the data, which is very suitable for flight time series data [4]. Figure 1 illustrates a single LSTM memory cell.

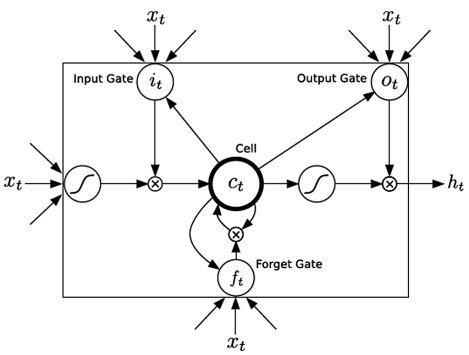

Fig. 1. Long Short Term Memory cell.

For the version of LSTM used in this paper, H is implemented by the following composite function:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t &= f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ h_t &= o_t tanh(c_t) \end{aligned} \tag{2}$$

where σ is the logistic sigmoid function, and i, f, o and c are respectively the input gate, forget gate, output gate, cell and cell input activation vectors, all of which are the same size as the hidden vector h. W_{hi} is the hidden-input gate matrix, W_{xo} is the input-output gate matrix etc. The weight matrices from the cell to gate vectors (f.g. W_{ci}) are diagonal. So element m in each gate vector only receives input from element m of the cell vector. The bias terms (which are added to i, f, c and o) have been omitted for clarity [10].

The original LSTM algorithm adopts a custom designed approximate gradient calculation that allows the weights to be updated after each time step. However, the full gradient can instead be calculated with back propagation through time.

3.4 Combination of Cluster and LSTM Analysis Model

After data preprocessing, we calculate the Spearman's rank correlation coefficient between each two series and get the correlation coefficient matrix. The heat map of this matrix is shown in Fig. 2:

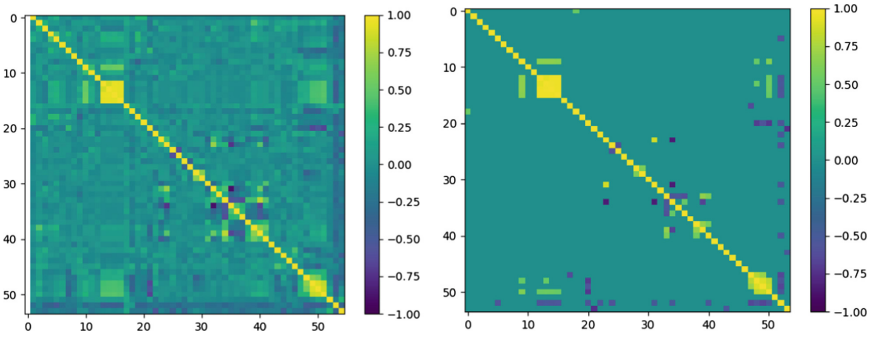

Fig. 2. The left figure is Spearman's rank correlation coefficient matrix heat map. The right figure is the Spearman's rank correlation coefficient matrix heat map which set the correlation coefficient that the absolute of it less than 0.5 be 0.

As is shown in Fig. 2, the brighter or darker point means that two series linked to this point is correlated. The brighter one shows the similar tendency of two series, while the darker ones shows the opposite but related tendency. In the process of clustering the series, our aim is gathering the series whose Spearman's rank correlation coefficient between them is closed to +1 or −1. Similar to the hierarchical group method in [11], this process is described in Table 1.

In the clustering process, we gather the high related series together. At the same time, we abandon the series which are not related to any other series.

After the preprocessing, we build an LSTM prediction model to analyse the data. Our LSTM model includes LSTM layers and full connected layers. We use 128 LSTM layers to get the information from input data and 19 dense layers to export the predicted output. The loss function as MAE (Mean Absolute Error) and the optimizer is Adam. The whole framework is shown in Fig. 3.

Table 1. Pseudo-code of clustering based on the Spearman's rho

Input Parameters: series set=$\{x_1, x_2, \ldots, x_m,\}$;
Similarity function: Correlation_Coefficient (a_i, a_j);
Threshold value:λ;
Aim number of cluster:k

for i = 1,2, … , m **do**
  $C_i = \{x_i\}$
  for j = 1,2, … , m **do**
    $r_{i,j}$ = Correlation_Coefficient (x_i, x_j);
  end for
end for
for i = 1,2, … , m **do**
  for j = 1,2, … , m **do**
    M(i,j) = $r_{i,j}$;
    M(j,i) = M(i,j)
  end for
end for
set original cluster number: q = m
while q > k **do**
  find two cluster C_{i^*} and C_{j^*} whose absolute value of Spearman's rho is biggest
    if Abs(M(i^*,j^*)) < λ **do**
      break whole process
    /* When all the distance between two clusters are
    less than the threshold, the similarity between the two
     clusters are small. */
    else do
      merge C_{i^*} and C_{j^*} : $C_{i^*} = C_{i^*} \cup C_{j^*}$
      for j = $j^* + 1, j^* + 1, \ldots, q$ **do**
        renumber the C_j to C_{j-1}
      end for
      delete the j^{*th} row and j^{*th} column
      for j = 1,2, … , $q - 1$ **do**
        for i in C_{i^*} **do**
          $r_{i,j}$ = Correlation_Coefficient (x_i, x_j)
        end for
        M(i^*, j)=max($\mathbf{r}_{*j}$)
      end for
    q = q − 1
end while

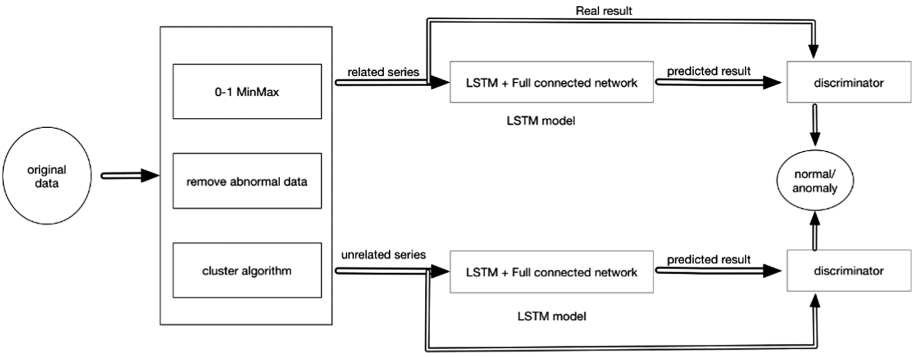

Fig. 3. Correlation coefficient based cluster and LSTM prediction model.

4 Experimental Results and Analysis

We set up our experiments to evaluate the effectiveness of our method for anomaly detection of time series. The settings are as follows.

The operating system we use is Ubuntu16.04. Our server has "Intel(R) Xron(R) CPU E5−2680 v4 2.40 GHz" CPU and "NVIDIA Tesla K80" GPU. The language we choose is python3.5 and the main toolkit we use is Keras, Tensorflow, matplotlib, numpy and pandas.

We use the real data of the COMAC C919 aircraft during a test flight. The data contain the time series of 56 sensors at a 6K sampling frequency. We deal with the data using the MinMaxScaler method. During the process, we find certain sample values to be constant. These abnormal data are eliminated in the preprocess. We extract the preprocessed data from 54 sensors at 10,000 sample points for further processing.

After calculating the correlation coefficient and the clustering algorithm is in operation, we get the required classes. Figure 4 shows two clusters with strong correlations obtained by clustering. For comparison, we also draw a sequence group with weak correlation (Fig. 5).

In the comparison of the two figures, it can be found that clustering results in a highly correlated sequence, which also has a visually related growth trend. Among them, the orange line in the second cluster (Fig. 4 bottom) means that the line has an opposite but very relevant trend with other sequences.

Before training, we change the format of data. We use the data of previous 10 sampling points to predict the data of later sampling point. Since each sampling point has 19 dimensions of data, our input is a 10×19 matrix and output is a 1×19 matrix.

In the next experiment, we design the LSTM model for training and predictive analysis. We put the closely related sequences and the sequences which are used in [1] into our LSTM model. After several rounds of training, we get the maps of MAE and training rounds, as shown in Fig. 6.

After error calculation, we get the following RMSE (Root Mean Square Error) in Table 2.

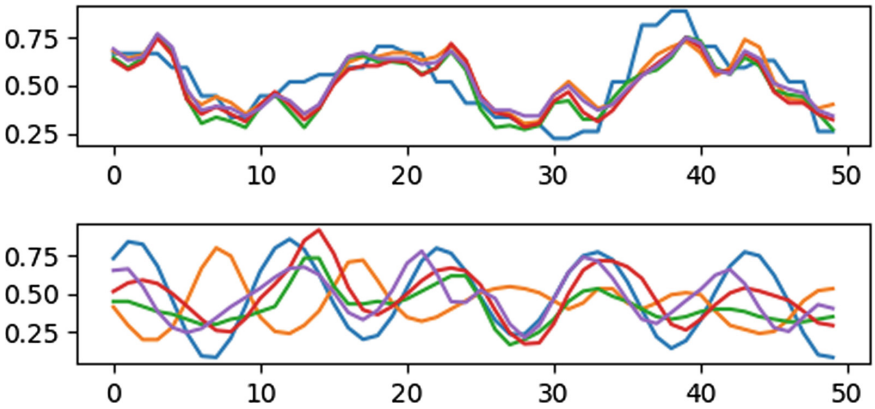

Fig. 4. Two clusters of related series.

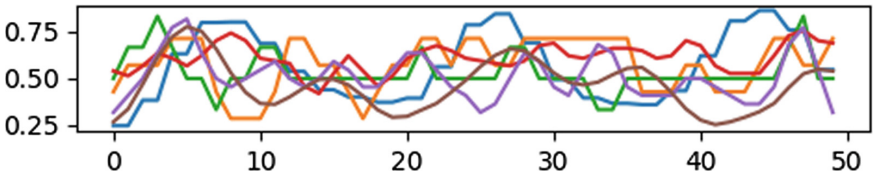

Fig. 5. Example of unrelated series.

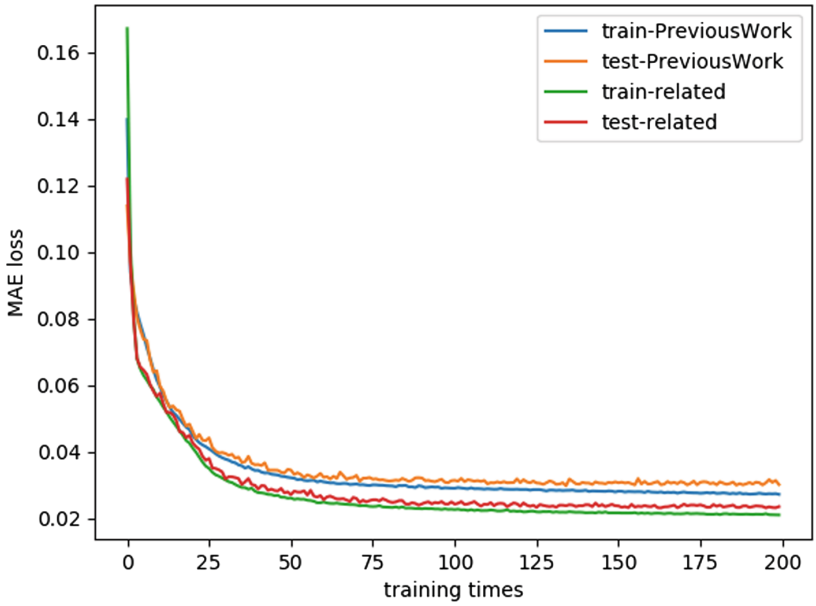

Fig. 6. MAE loss of related series and previous used series during the LSTM training.

Table 2. RMSE of the previous model and our model

	Training dataset	Test dataset
Series in reference [1]	0.002232	0.002378
Series selected via our clustering method	0.001731	0.001790

Through comparative experiments, we find that the more relevant data sets have faster convergence rate and less loss of convergence results than the randomly selected series used in previous work. Models based on clustering and LSTM have better performance in high latitude time series analysis.

5 Conclusion

We propose a novel method of clustering and predicting multidimensional aircraft time series whose analyses are challenging in data science. Given COMAC's C919 flight test data, we observe that uncorrelated information and data redundancy in high latitude sequences can interfere with the analytical capabilities of the LSTM based prediction model for the time series of flight test data. With these observations, our method integrates clustering with an LSTM model to select time series with high correlation from high latitude sequences, which improves the accuracy of the LSTM prediction model compared with recent work. Our research can be further extended to other scenarios of time series data analyses.

Acknowledgment. This work is partially supported by National Key Research & Development Program of China (2017YFA0206104), Shanghai Municipal Science and Technology Commission and Commercial Aircraft Corporation of China, Ltd. (COMAC) (175111105000), Shanghai Municipal Science and Technology Commission (18511111302, 18511103502), Key Foreign Cooperation Projects of Bureau of International Co-operation Chinese Academy of Sciences (184131KYSB20160018) and UK EPSRC (EP/L016796/1, EP/N031768/1 and EP/P010040/1).

References

1. Cao, Z., Zhu, Y., et al.: Improving prediction accuracy in LSTM network model for aircraft testing flight data. In: IEEE International Conference on Smart Cloud (2018)
2. Hsu, H., Hsieh, C.: Feature selection via correlation coefficient clustering. J. Softw. **5**(12), 1371–1377 (2010)
3. Gauthier, T.: Detecting trends using spearman's rank correlation coefficient. Environ. Forensics **2**, 359–362 (2001)
4. Nanduri, A., Sherry, L.: Anomaly detection in aircraft data using recurrent neural networks. In: Integrated Communications Navigation and Surveillance (ICNS) Conference (2016)
5. Grabusts, P., Borisov, A.: Clustering methodology for time series mining. Sci. J. Riga Tech. Univ. **40**(1), 81–86 (2009)
6. Singhal, A., Seborg, D.: Clustering multivariate time-series data. J. Chemom. **19**, 427–438 (2005)

7. Funie, A.-I., Grigoras, P., Burovskiy, P., Luk, W., Salmon, M.: Run-time reconfigurable acceleration for genetic programming fitness evaluation in trading strategies. J. Signal Process. Sys. **90**(1), 39–52 (2018)
8. Gai, K., Qiu, M., Zhao, H., et al.: Dynamic energy-aware cloudlet-based mobile cloud computing model for green computing. J. Netw. Comput. Appl. **59**, 46–54 (2016)
9. Bara, A., Niu, X., Luk, W.: A dataflow system for anomaly detection analysis. In: International Conference on Field Programmable Technology (2014)
10. Graves, A.: Generating sequences with recurrent neural networks. https://arxiv.org/abs/1308.0850
11. Cui, L., Luo, Y., Li, G., Lu, N.: Artificial bee colony algorithm with hierarchical groups for global numerical optimization. In: Qiu, M. (ed.) SmartCom 2016. LNCS, vol. 10135, pp. 72–85. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52015-5_8
12. Gai, K., Qiu, M., Liu, M., Zhao, H.: Smart resource allocation using reinforcement learning in content-centric cyber-physical systems. In: Qiu, M. (ed.) SmartCom 2017. LNCS, vol. 10699, pp. 39–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73830-7_5