上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS

论文题目：电厂运行数据数字化应用前处理系统研究

学生姓名：刘子屹

学生学号：5140219070

专　　业：新能源科学与工程

指导教师：忻建华

学院(系)：机械与动力工程

DATA PREPROCESSING FOR STEAM TURBINES

At present, the digitalization of power plants is appealing to more and more attention while one of the core topics is the knowledge discovery from the operation data. But the related data preprocessing methods is rarely mentioned in the literature, especially in China. Millions of sensor data are generated everyday in the power plant, however, due to the instability of sensors and interference during data transmission, the collected raw data is usually contaminated. The low quality of data will seriously affect the outcomes of data mining. Therefore, the study on data preprocessing for steam turbines is indispensable.

The thesis first illustrate the scope of data preprocessing systematically including two main processes, i.e. data preparation and data reduction. Specifically, the data preparation section covers data integration, redundancy removing, data cleaning and data transformation, while the data reduction sections includes feature selection, instance selection and discretization. Furthermore, the data cleaning process can be divided into anomaly detection and data imputation. Especially, a detailed literature review for anomaly detection is constructed from method selection to method classification and comparison. Statistical, Nearest neighbor-based, Clustering-based and Classification based techniques are four common methods for anomaly detection. Considering the industrial needs and research scope, the study focuses on three key problems: data redundancy, anomaly detection and data imputation.

With real operating data of the steam turbine, the study employs and compares various methods to solve the aforementioned problems. In terms of the high cost of manual data labeling, unsupervised methods are given priority. Provided by Shanghai Electric, the raw data is continuously three-month minutely data with 66,942 instances and 15 attributes. The attributes set is comprised of 1 power, 4 valve position feedback, 7 temperature, 13 pressure attributes. The complete data preprocessing is constructed with five steps:

i. Read and classify sensor data of steam turbine into multiple redundant sensor data sets and other measuring point data set, and the data sets respectively enter similar process as follows
ii. Determine whether it can be converted to floating-point data (i.e., determine whether it is numeric data), so that to detect non-numeric data (including missing data, string, etc.), and mark it as None, and record its location in the non-numeric data location set.
iii. Perform the first simple forward interpolation for the None data in step 2 as to meet the requirements of the common anomaly detection algorithm, so all the data sets with only numerical data enter the anomaly detector.
iv. Determine whether it is an anomaly, record its location in the anomaly data location set, and then the combine the anomaly data location set with the non-numeric data location set in step 2 to obtain the data location set to be replaced.
v. For the non-redundant sensor data set, the data imputation model performs prediction on the location to be replaced. The results merges with the original normal data to obtain a clean data set. As for the redundant sensor data sets, if the attributes at any time instance

are all abnormal, the above prediction is directly applied as well. Otherwise, the proposed method “nearest average of the normal data” is utilized to reduce the redundant sensor data set into a clean single-dimension data set.

Overall, the whole preprocessing lays a foundation for the subsequent data applications such as condition monitoring and fault diagnosis.

As for data redundancy, the Pearson's Correlation Coefficient is calculated for correlation analysis, and the "nearest" average value method is defined for the sensor redundancy. The results of correlation analysis can provide insights for the feature selection of subsequent KDD processes, besides, the results can support the attribute grouping for the experiments of multi-dimensional anomaly detection. However, since this study focuses on the data preprocessing, it is necessary to ensure the integrity of the data attributes in the initial preparation stage of KDD. Therefore, the removing of the redundant sensor data sets becomes the focus of the study. The basic principle of the "nearest" average value method is to select the most similar normal sensors at any one time instance and calculate the average value as the only output data at that moment. The so-called "nearest" average value, that is, under the condition that the given standard deviation threshold is satisfied and the least normal sensors are eliminated, the simple arithmetic mean of the normal sensor group with the smallest standard deviation.

With regard to anomaly detection, the experimental mechanism for artificially introducing anomalies is constructed to test the statistical Z-score, the nearest neighbor based Local Outlier Factor and the unsupervised-learning Isolation Forest methods. Besides, the relationship between the dimensionality of the data set and the detection results is also analyzed. The results illustrate that:

a) The optimal anomaly threshold is $\mu \pm 3\sigma$ when both anomaly detection accuracy and fault rate are in a relatively acceptable range. In this case, the average accuracy $\eta = 71.8\%$, the average fault is $\theta = 29.4\%$ with the average detection time $T_{test} = 0.0046s$ per 66,942 instances. Besides, the results illustrate that the Z-score method is sensitive to the value with high degree of abnormality, especially when the degree of abnormality is higher than ±80%, the accuracy is above 95%. Besides, the Z-score methods performs better on temperature attributes ($\eta > 85\%$) than pressure and power attributes.

b) As for Local Outlier Factor, the results of multi-dimensional detection is significantly better than that of single-dimensional detection; under specific anomaly scenario ($\mathrm{threshold} = 80, \mathrm{ratio} = 0.5$), the detection accuracy is continuously improved with the increase of the neighbor parameter k. As the growth rate slows down, the accuracy gradually reaches a peak value of about 75% at which $k \approx 500$; and the detection time and k exhibit a near-exponential positive correlation trend, which reflects the high complexity of the nearest neighbor-based methods. What’s more, the accuracy of the high correlation attribute combinations 4 and 5 is relatively higher, so it can be inferred that the LOF is more suitable for the highly correlated multi-dimensional data set.

c) The anomaly detection accuracy of the Isolated Forest is continuously improved with the increase of the abnormal degree and the abnormal proportion. At the same time, the influence between the interval of low abnormal degree is more significant, and the abnormal proportion has a roughly linear relationship with the detection accuracy, which indicates that Isolated Forest is more suitable for data set with abundant anomalies.

Meanwhile the abnormal degree is as high as ±80%, the method has outstanding anomaly detection performance (more than 90%), even in the case of a lower abnormal proportion (such as 1%). Similarly, when multi-dimensional data is used for detection, the detection performance gradually deteriorates with the increase of the dimension.

Concerning data imputation, prediction models including the Exponential Moving Average (EMA) and Autoregressive Integrated Moving Average (ARIMA) are developed to implement a complete process of data cleaning:

a) As for the EMA model, the prediction results of the full time series of the 15 attributes indicate an average RMSE of 0.732, deviation ratio of 8.98E-03, and prediction time of 0.006 seconds per instance. The smoothing constant can automatically be optimized through the simulations. However, the prediction accuracy depends heavily on the size of the training sample and the trend of the time series, which results in a large deviation for individual instances.

b) ARIMA can achieve extremely high fitness and prediction accuracy with appropriate parameters (p, d, q) by observing the data trend, ACF map, and PACF map after differencing for specific time series. The example instance demonstrates high performance with a training RMSE of 0.218, predicting RMSE of 0.207, and deviation ratio of 4.24E-04, but the prediction time per instance is nearly 60 times longer than that of EMA, so this method is more suitable for offline data prediction.

In summary, this research not only provides practical solutions for real issues of power plants, but also lay a reference foundation for academic researchers in the field of data mining. Nowadays, digitalization is an inevitable trend in the industrial development. If the data is compared to the fuel that drives the power plants smarter, the data pre-processing system can be regarded as the catalyst for improving fuel efficiency. In the era of big data, the traditional energy industry must realize digital transformation through the integration of advanced technology portfolio and management strategies, so that to remain competitive and dynamic. From smart devices to intelligent control, to intelligent production monitoring and to intelligent management, the realization of the Digital Power Plant has a long way to proceed, but there is no doubt that challenges and opportunities coexist along the journey.