IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 10, 2017, accepted February 1, 2018, date of publication February 14, 2018, date of current version March 13, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2805908

Multi-View Stacking Ensemble for Power Consumption Anomaly Detection in the Context of Industrial Internet of Things

ZHIYOU OUYANG[1,2], XIAOKUI SUN[1,2], JINGANG CHEN[3], DONG YUE[1,2,4], (Senior Member, IEEE), AND TENGFEI ZHANG[2], (Member, IEEE)

[1]Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2]School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[3]School of Economics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[4]Jiangsu Engineering Laboratory of Big Data Analysis and Control for Active Distribution Network, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Corresponding author: Dong Yue (medongy@vip.163.com)

This work was supported in part by the National Natural Science Key Fund under Grant 61533010 and in part by the Primary Research & Development Plan of Jiangsu Province under Grant BE2016184.

ABSTRACT Anomaly detection of power consumption, mainly including electricity stealing and unexpected power energy loss, has been one of the essential routine works in power system management and maintenance. With the help of Industrial Internet of Things technologies, power consumption data was aggregated from distributed various power devices. Hence, the power consumption anomaly was able to be detected by machine learning algorithms. In this paper, a three-stage multi-view stacking ensemble (TMSE) machine learning model based on hierarchical time series feature extraction (HTSF) methods are proposed to solve the anomaly detection problem: HTSF is a novel systematic time series feature engineering method to represent the given data numerically and as input data for machine learning algorithms, while TMSE is designed to ensemble meta-models to archive more accurate performance by using multi-view stacking ensemble method. Performance evaluation in real-world data shows that the proposed method outperforms the existing time series feature extraction means and dramatically decreases the time consumed for ensemble learning process.

INDEX TERMS Internet of Things, machine learning, smart grids, time series analysis, feature extraction, power consumption, anomaly detection.

I. INTRODUCTION

Detection of abnormal electricity behaviors has plagued power operating companies in China because of numerous power system faults and profit reduction caused by the abnormal electrical consumption. Therefore, accurate detection of abnormal behaviors has long been of concern among power supply enterprise. The primary methods of electricity stealing and leakage include power anomalies, unusual load, line loss abnormalities in [1]. It is essential to make accurate detection of abnormal behaviors to reduce the risk of power companies and standardize the user's response. With the help of Industrial Internet of Things (IIoT), Advanced Metering Infrastructure (AMI) was deployed in power systems and aggregating information from distribution power systems, as shown in the following Fig.1:

As shown in the above IIoT based power consumption anomaly detection architecture, the power consumption data of electricity meter, mainly including internet of things based wireless electricity sensor and RS232/RS485 connected electric quantity acquisition instrument, are collected and sent to the cloud-based data center. Hence the user power consumption behaviors can be analyzed by machine learning algorithms.

The usual method of dealing with the anomaly detection was building an outline of typical instances and then identify instances that do not fit the conventional profile as anomalies in [2]. In [16], normal user behaviors are used to train a regular model by the support vector machines (SVM) algorithm. The unknown data are compared with the normal model, then the low similarity is considered as abnormal behavior.

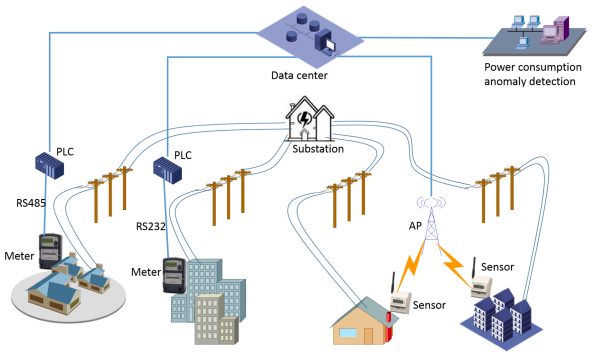


FIGURE 1. IIoT based power consumption anomaly detection architecture.

Statistical methods, classification-based methods and clustering-based methods are also used to solve the problem. Among them, the statistical method is a traditional method. For big data sets, in [3], it gives accurate results by dealing with two critical applications. In [4], it thought that classification algorithm is used to outlier detection, which leads to two issues: one is the lack of labeled outlier data, and another is the relatively high computational requirement. In [5], the clustering algorithm is used to cluster local outlier. In [6], it demonstrates a way of performing time series anomaly detection via generated states and rules and introduces an algorithm named Gecko for clustering time series data.

In many studies, there often encounter a problem that only rarely labeled data was available even if a significant amount of un-labeled data has aggregated. For example, in this paper, only about 100,000 labeled users while there are hundreds of millions of un-labeled users. On the other hand, IIoT based AMI sensors can obtain numerous properties that involved in power consumption and user behaviors, but many of the labeled user data only including user registration, daily power consumption, which limited the use of machine learning algorithms. One of the effective approaches to solving those problems is multi-view learning technology. In this paper, an anomaly detection model for power consumption based on hierarchical time series feature extraction with multi-view learning based ensemble learning solution is proposed, which improves the performance of detection accuracy with only limited data used. The features of this model are as follows:

- Limited power consumption data needed.
- Systematical feature engineering proposed.
- Outperform existing time series feature engineering.
- Improved multi-view stacking.
- Won the 3rd prize in 2016 CCF DBCI.

To this end, the major contribution of our work is summarized follow a data-driven low-cost solution using only limited data is proposed to avoid any extended AMI hardware. This paper uses hierarchical time series feature extraction with supervised binary classification to build a model that performs well.

The rest of this paper is organized as follows: firstly, the existing approaches for time series feature extraction and multi-view stacking ensemble methods are reviewed, then, the proposed hierarchical time series feature extraction method and three-stage multi-view stacking ensemble model is described in detail. Moreover, the data that are real-world power consumption data of over 50000 customers that offered by State Grid of China are used. Perform experiments on the proposed feature engineering method with proposed three-stage multi-view stacking ensemble has compared to existing time series feature extracting methods and supervised machine learning algorithms, which shows that the proposed method has better accuracy.

II. RELATED WORKS

A. ANOMALY DETECTION METHODS

Time series is a numerical value of some statistical indicators, arranged in chronological order formed by the sequence. It is used in many fields. References [7] and [8] put forward and analyzed the combination of spectral decomposition and feature selection method for discrete cosine transform (DCT) and discrete wavelet transform (DWT) in the field of the time series classification problem. In [9], time series algorithm establishes the prediction model to use for clinical decision support by integrating information from multiple characteristics of electrocardiogram (ECG). In [10], it proposed a feature extraction procedure for a Brain-Computer Interface (BCI) application, and the feature is extracted from electroencephalogram (EEG) features from time series method performs the right and left movement image. In this paper, the time series method is used for detection of abnormal behaviors of power users. A time series model will be used, like equation (1). $x(t)$ and $O(t)$ represents respectively the input and output. The pulse function of the input to the output is denoted by $f_i(t)$ in [11].

$$\begin{cases} O(t_1) = x(t_1) * f_1(t) + d_1(t) \\ O(t_2) = x(t_2) * f_2(t) + d_2(t) \\ \cdots \\ O(t_n) = x(t_n) * f_n(t) + d_n(t) \end{cases} \tag{1}$$

where $*$ represents convolution, the additional noise is $d_i(t)$ and $x(t_1), x(t_2), \cdots, x(t_n)$ can be repressed as:

$$x(t_1) \to x(t_2) \to \cdots \to x(t_v) \to \cdots \to x(t_n) \tag{2}$$

where $t_{v+1} = t_v + \delta$. As long as the $f_i(t)$ of each transmission channel is obtained, the outlier can be detected through the system characteristics of the parameters. $d_i(t)$ can be obtained by the time series model. In [11], a time series feature extraction algorithm was proposed, which classifies or regressions the available features in the early stage of machine learning, and uses feature importance filter to combine established feature extraction methods. The data processing tiers of [11] are illustrated in the following fig.2:

A systematic but imperfect time series feature extraction method called hierarchical time series feature extraction methods is proposed in [12]. However, the principle behind

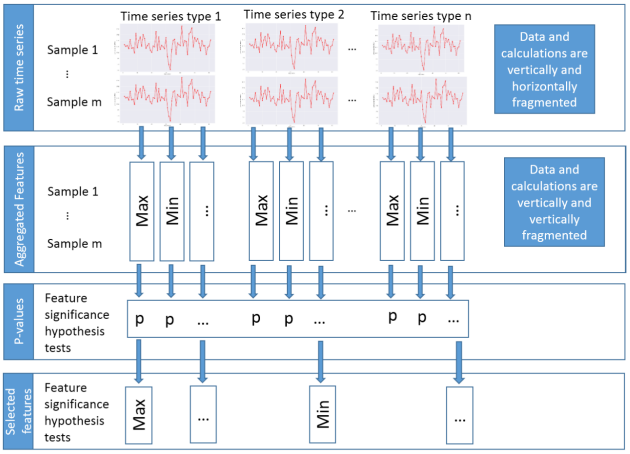


FIGURE 2. Data processing tiers of tsfresh.

the proposed feature engineering method and mathematics or statistic equations were not introduced clear enough, as well as the multi-view stacking ensemble method that based on the proposed feature engineering to improve performance. Hence, in this paper, the hierarchical time series feature extraction methods are redefined with more mathematic equations or statistic principles, and the ensemble model using those features to construct more effective anomaly detection machine learning model for power consumption is also be introduced.

B. MULTI-VIEW STACKING

Multi-views stacking is a combined algorithm based on Multi-views learning and stacked generalization (sometimes called stacking), which is an enhanced learning method which can ensemble several learners to obtain better results. The multi-views learning, stacked generalization, and multi-view stacking are introduced in the following, separately.

1) MULTI-VIEW LEARNING

It is difficult to recognize its essence of objects from a certain aspect or point of view because that most objects are multidimensional: For a web-page, it can be characterized not only by the text but also by the hyperlinks pointing to the page. For a video, it could be represented by the image in its content and audio accompanied by it.

The current theories on multi-view learning can be classified into three categories: 1) Canonical correlation analysis (CCA) that was proposed in [13] and works on a paired data set to find linear transformations; 2) Correlations are maximized for one view; 3) Co-training algorithm was introduced in [14] for semi-supervised classification. Take the CCA (e.g., data represented by two views) as an example, suppose the data set $[(x_1, y_1), \ldots, (x_n, y_n)]$ has two views, one is $x = (x_1, x_2, \ldots, x_n)$ while the other is $y = (y_1, y_2, \ldots, y_n)$, hence the purpose of CCA is to find two projection directions and to maximize the following linear correlation coefficient.

In this work, many features extracted from dataset will be regarded as the views which can be used in multi-view stacking algorithm.

2) STACKED GENERALIZATION

Stacked generalization introduced in [15] is an ensemble method for combining multiple learners. It was used for classification and surface-fitting. The features from multi-views learning can be trained by many algorithms and then get the final results by stacked generalization. The overall procedure comprises the following steps: (1) other algorithms are used to train the existing data and get several learners; (2) an algorithm is trained to use all the predictions of the other algorithms as additional inputs for the final prediction by stacked generalization.

3) MULTI-VIEWS STACKING

Multi-views stacking method consists of training one first-level learner for each view and combining their outputs using stacked generalization in [16]. Multi-views stacking algorithms are often used to extend features when multiple algorithms are used to train the same data set. In this paper, summary features, shift features, transform features and decompose features which represent a feature group respectively make up the four views of the raw data. Then, the basic models will be taken to train the four views features and obtain meta-learners of each view. Each meta-learner can give an output for each view. These outputs are made up of predicted labels and associated prediction probabilities for each of the k classes. Thus, the final input features are obtained. The final features will be used to train final model. The general steps are as follows:

step1: divide the data set into k views: $(V_1, V_2, \ldots, V_k)$;
step2: the basic algorithms are used to train models for each view and the output probabilities of the first-level learners are averaged, referred to as $(p_1, p_2, \ldots, p_k)$. $(l_1, l_2, \ldots, l_k)$ are the predicted labels of each first-level model;
step3: the combined feature vectors are $(l_1, l_2, \ldots, l_k, p_1, p_2, \ldots, p_k, y)$, where y is the true label;
step4: use the combined feature vectors to train new model;

III. PROPOSED METHODS

A. HIERARCHICAL TIME SERIES FEATURE EXTRACTION

To use the power consumption data to detect abnormal energy consumption activities, one of the best methods is to treat it as a time series classification problem by using supervised classification algorithms. Therefore, the key to solving the problem is to extract the time series features. The extracted time-series features are mainly used to find the abnormal sample distribution rules as well as the information from normal power consumption activities. In this section, a hierarchical time series feature extraction method that extracts the features of the power consumption time series systematically

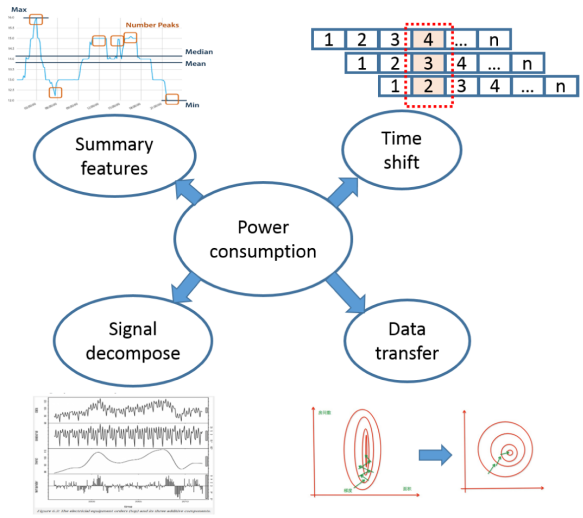


FIGURE 3. Hierarchical time series feature extraction structure.

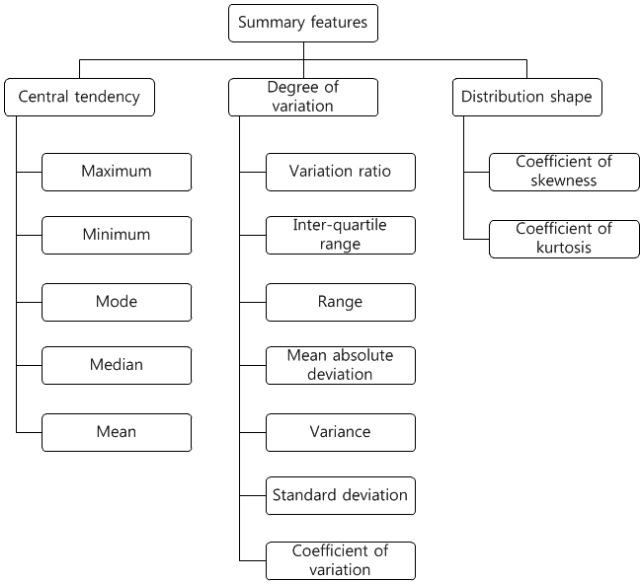


FIGURE 4. Summary features structure chart.

is presented. The features are consisted of summary features, shift features, transform features and decompose features, as shown in the following fig.3.

1) SUMMARY FEATURES

Summary features are statistical variables that describe the distribution of samples over various time periods, also called time-window statistical variables. Summary features consist of Central tendency, Degree of variation, Distribution shape, such as maximum value, minimum value, median, average value, and variance. These features are similar since all the features are extracted from a time-window segment of the time series. However, the time-window segment refers to a set of time series data separated by the start time and the end time. The principle of feature extraction of summary features is shown in fig.4.

Some symbolic features are defined as follows, in which X indicates a time-window $[t_1, t_n]$,and x_{t_i} denotes power consumption value at time t_i:

a: CENTRAL TENDENCY

It refers to the extent to which set of data is closer to a central value, which reflects the location of a set of data center points.

maximum: The largest value in a set of data as defined in the following equation:

$$\max x_{t_i} \tag{3}$$

minimum: The smallest value in a set of data as defined in the following equation:

$$\min x_{t_i} \tag{4}$$

mode: The highest frequency of occurrence in a group of data;

median: The middle order of a group of data as defined in the following equation:

$$median = \begin{cases} x_{t_{(n+1)/2}} & : t_n \ is \ uneven \\ 1/2 * (x_{t_{n/2}} + x_{t_{n/2+1}}) & : t_n \ is \ even \end{cases} \tag{5}$$

mean: statistic variable that describes the degree of data concentration as defined in the following equation:

$$\tilde{x} = \frac{1}{n} \sum_{i=1}^{n} x_{t_i} \tag{6}$$

b: DEGREE OF VARIATION

The degree of variation is another important feature of the data distribution, which reflects the distance between the values of variables and their center values. The greater the degree of variation, the worse the concentration of the measured values of the set of data; the smaller the degree of variation, the better the representation.

Variation ratio: The ratio between the total number of non-public and the total number of them:

$$V_r = 1 - \frac{f_m}{\Sigma f_i} \tag{7}$$

Interquartile range: The difference between the upper quartile and the lower quartile as defined in the following equation:

$$Q_d = Q_u - Q_L \tag{8}$$

Range: The data obtained by subtracting the minimum value from the maximum value of a group of data as defined in the following equation:

$$R = \max x_{t_i} - \min x_{t_i} \tag{9}$$

Mean absolute deviation: The arithmetic mean of the absolute values of the deviations of all units from their arithmetic mean as defined in the following equation:

$$M_d = \frac{\sum_{i=1}^{n} |x_{t_i} - \bar{x}|}{n} \tag{10}$$

Variance: The average of the squared differences between the mean of each sample value and the overall sample value as defined in the following equation:

$$s^2 = \frac{\sum_{i=1}^{n}(x_{t_i} - \bar{x})}{n-1} \tag{11}$$

Standard deviation: The arithmetic square root of variance as defined in the following equation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_{t_i} - \bar{x})}{n-1}} \tag{12}$$

The coefficient of variation: The standard deviation and the average ratio, measure the degree of variation of the relative statistics as defined in the following equation:

$$v_s = \frac{s}{\bar{x}} \tag{13}$$

c: DISTRIBUTION SHAPE

Central tendency and the degree of variation are two essential features of data distribution, which can help to fully understand the characteristics of the data distribution: whether the shape of the data is symmetry, nor the distribution is flatness, as well as the degree of deflection. The coefficient of skewness and the coefficient of kurtosis is metric of the distribution shape.

The coefficient of skewness: The ratio of the mean to the median to the standard deviation measures the degree of skewness as defined in the following equation:

$$SK = \frac{n\Sigma(x_{t_i} - \bar{x})^3}{(n-1)(n-2)s^3} \tag{14}$$

The coefficient of kurtosis: The measure of flat or spike distribution of data as defined in the following equation:

$$K = \frac{n(n+1)\Sigma(x_{t_i} - \bar{x})^4 - 3[\Sigma(x_{t_i} - \bar{x})^2]^2(n-1)}{(n-1)(n-2)(n-3)s^4} \tag{15}$$

d: DECOMPOSITION FEATURES

It refers to the decomposition of power consumption time series data extracted features, and each decomposition feature represents a category portrait.According to the time series decomposition theory, the time series can be decomposed into four factors: trend (T), season (S), the special date (D) and random fluctuation (I). In other words, the time series can be fitted with a function of these four factors as the following equation:

$$x_t = f(T_t, S_t, D_t, I_t) \tag{16}$$

Based on the idea of factorization, the main purpose of extracting decompose features is to overcome the interference of other factors and to measure the influence of certain determinants on time series. It can found that the dataset contains the power data of January, March, April, and May, and the observation period is not long enough to fully reflect the influence of seasonal factors. Therefore, the seasonality factor could be removed.

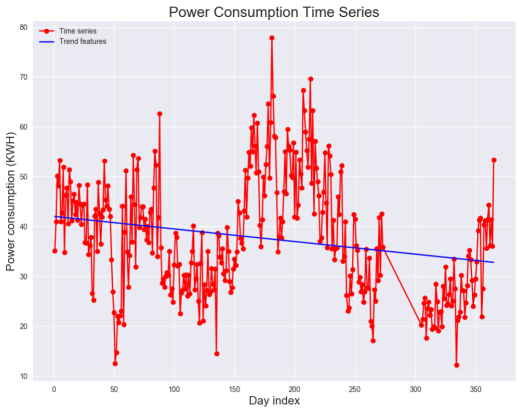


FIGURE 5. Trend features of power consumption time series.

After preliminary data exploration, it can be found that the use of electricity data is not subject to certain special dates such as the Spring Festival. Therefore, the special date factor can be carefully omitted. In summary, through the above analysis, the task is to extract the two features of the trend (T) and the random fluctuation (I).

Firstly, establish a trend regression process using equation (17):

$$x_t = c + \beta t + \varepsilon_t \tag{17}$$

Secondly, the trend features in the sequence were extracted by fitting a linear model as equation (18):

$$\hat{x_t} = \hat{c} + \hat{\beta} t \tag{18}$$

Thirdly, by subtracting the trend eigenvalue equation from the original sequence, the stochastic volatility feature could be obtained by equation (19):

$$\epsilon_t = x_t - (\hat{c} + \hat{\beta} t) \tag{19}$$

Hence, the decomposition of the time series that extracting the decomposition features of trend and random fluctuation was completed, and the effect diagrams are illustrated in following fig.5 and fig.6.

2) TRANSFORM FEATURES

a: LOGARITHMIC TRANSFORM

The two important attributes of continuous variable distribution are central tendency and degree of variation. According to the central limit theorem, it is reasonable to assume that the power consumption of a user obeys normal distribution. In other words, this means that although the user's electricity consumption is variable, they should be well distributed around an intermediate value. However, there is no exclusion of the emergence of outliers, and they may be caused by stealing power. The median and interquartile range are more robust in the presence of outliers than the more common mean

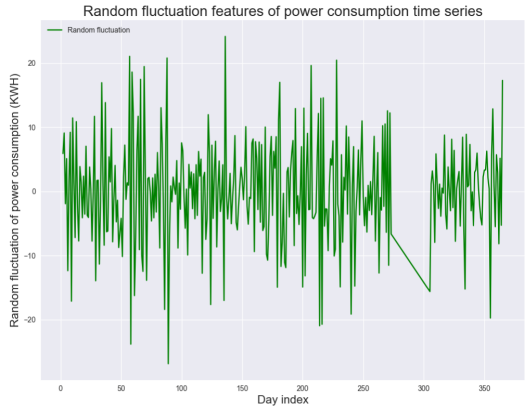


FIGURE 6. Random fluctuation features of power consumption time series.

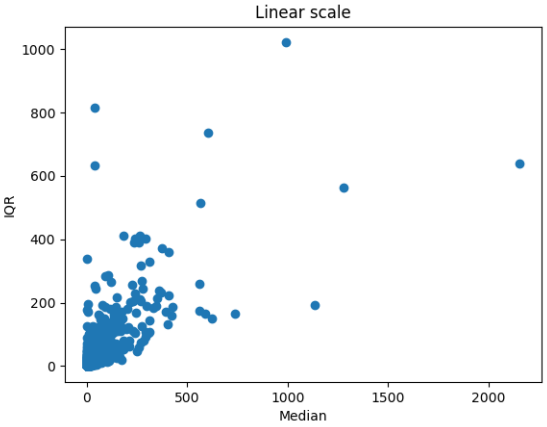


FIGURE 7. Linear scale diagram of median and quartile.

and standard deviations. So this paragraph adopts the median and interquartile range of the two statistics.

Calculate the median and quartile based on each user's electricity consumption data. They are drawn in fig.7.

Most of the data can be found in a more concentrated distribution. However, a few points are obviously far away from other points. Therefore, it is hard to read in fig.6 and is not conducive to extracting the distribution characteristics of most of the data. Therefore, the problem can be solved by using the logarithmic function as defined by the following equation:

$$F = \ln(x + 1) \tag{20}$$

Obviously, the distribution of data in fig.7 becomes more readable, with more concentrated data distribution and fewer outliers. Another noteworthy rule is that in either fig.7 or fig.8, the median and quadrant distances for many users are roughly along a diagonal line, demonstrating that users have similar power distribution in conclusion.

b: FOURIER TRANSFORM

Frequency domain analysis is to transform the time-domain data x_t into the frequency domain data x_f by using the Fourier

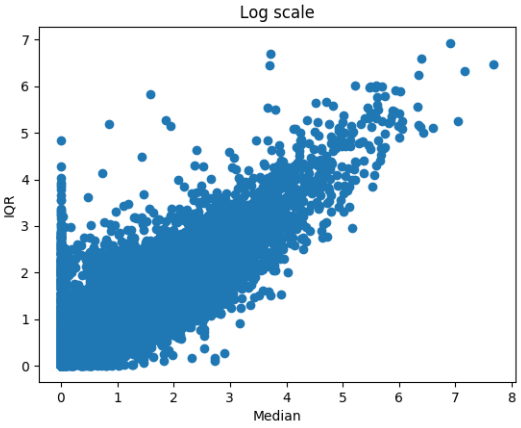


FIGURE 8. Log scale diagram of median and quartile.

transform, to dig out the features of the data from another perspective. Time-domain analysis can only reflect the signal amplitude changes over time, and it is difficult to reveal the signal frequency components and the size of the frequency components. Therefore, the Fourier transform method was used for frequency analysis of electricity consumption data to dig out further the data-related information, which can be defined by the following equation:

$$ff_t X_t = \sum_{i=t_1}^{t_n} x_t exp(-\frac{2\pi ik(i-1)}{n}) \tag{21}$$

3) SHIFT FEATURES

Shift features are features of k-step differential extraction of the time series of electricity consumption, including decrement value, decreasing rate, continuous decreasing times, etc. Shift features are used to describe the changing characteristics of the time series of electricity consumption, which is also expected to detect abnormal changes in electricity consumption. For example, abnormal electricity usage such as stolen electricity results in a significant reduction in electricity usage and this reduction can persist for some time. Therefore, it is necessary to excavate the time-series data of electricity consumption, and decrement value, decreasing rate, continuous decreasing times and other indicators that can provide relevant information. Once the above indicator reaches a threshold, It was reasonable to suspect that the user has stolen electricity and then conduct further analysis of the user. Compared with direct analysis, this preliminary mining can save resources and improve the recognition accuracy.

Related indicators are defined as follows:

- k-step difference: The subtraction between two sequence values separated by k is called a k-step difference operation

$$\Delta_k x_t = x_t - x_{t-k} \tag{22}$$

- decrement value: The one-step difference between the summary feature of the time-window t and the

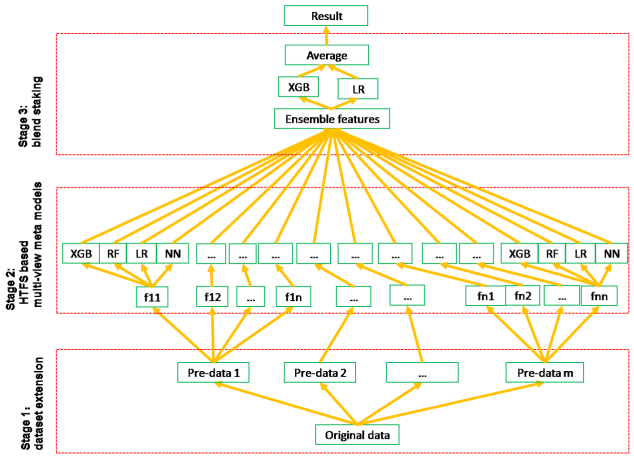


FIGURE 9. Three-stage multi-view stacking.

time-window t-1

$$\Delta_1 x_t = x_t - x_{t-1} \tag{23}$$

- decreasing rate: The amount of change in the decrement value of the time window t and the time window t-1

$$DR_t = \frac{\Delta_1 x_t - \Delta_1 x_{t-1}}{\Delta_1 x_t} \tag{24}$$

- continuous decreasing times: The number of which decrement value continues to be negative.

B. THREE-STAGE MULTI-VIEW STACKING ENSEMBLE

Ensemble learning is one of the most effective methods to improve machine learning performance, so as for anomaly detection machine learning models in which a small accuracy improvement may lead to substantial economic profits. However, the existing multi-view stacking ensemble methods required much more calculating resources and time. In this section, the three-stage multi-view stacking ensemble method that designed to improve anomaly detection accuracy without a large scale of calculating resource incremental is proposed based on the multi-view learning and stacked generalization consisting of training a model from each feature set extracted by HTSF and combining them with stacking.

The proposed three-stage multi-view stacking ensemble method consists of three stages: data extent stage, HTSF based multi-view meta-models and stacking, as shown in the following fig.9. The first stage that named data extent stage uses different missing-data processing methods to generate different datasets so that the potential information under missing-data can be represented from several views. Taking the power consumption data as an example, missing-data can be taken as abnormal activity and fill with numeric value -999 that indicates NAN value for decision tree models as pre-data 1, or fill with latest non-missing value as pre-data 2. Experiments show that different feature engineering methods have different performance on those extended data sets, which is essential for ensemble learning.

TABLE 1. Historical power consumption data.

Column name	Format	Sample
user_id	string	1234567890
day_index	int	1
meter_value	float	12.34

The second stage is HTSF based on multi-view meta-models that extracting features from datasets generated by the first stage using HTSF methods and building meta-learner due to stacking theory, as shown in Fig.3, using the extracted features. Several supervised learning algorithms like xgboost(XGB), Random Forest(RF), Logistic Regression(LR) and Neural Networks(NN) are used to build meta-learners for each view and combine their outputs (stealing probability range from 0 to 1) as feature data of the third stage. In the power consumption anomaly detection case, one view will comprise the information coming from one of the pre-data using one of the hierarchical time series feature extracting methods.

The third stage is stacking that regards predicted power stealing probabilities of meta-models as input features sets and uses xgboost and logistic regression to train and predict the stealing probability separately, and then calculate the average the two stealing probabilities for each customer as the final prediction.

IV. PERFORMANCE EVALUATION

The proposed HTSF based three-stage multi-view stacking model is implemented and analyzed in over 54,000 customers' historical power consumption for anomaly detection with a comparison to existing time series feature extraction methods tsfresh. The formation of user historical power consumption data is defined in table 1, in which the user_id is a unique id encoded by location and user type, the day_index indicates daytime that indexed from 01/01/2014 and meter_value is user daily power consumption values (KWH).

It was popular that there are missing data in historical power consumption data, as shown in the following fig.10, because of device failure, communication packet losses or power line topology change and so on. Moreover, the missing data also contains some information about the abnormal power consumption because many of the electricity stealing methods are to destroy the communication of electricity and stop the power consumption data probing. In this case, the missing data of power consumption data can be filled with 0 or -1 that reduce human interventions and regard as pre-data 1 and pre-data 2. On the other hand, because the missing data have different dates and should be filled using Linear Interpolation Method (LIM) for neural networks and regard as pre-data 3.

To illustrate the affection of missing data filling methods for different machine learning algorithms, feature engineering using tsfresh with minimum features configuration (named tsfresh-mini) and tsfresh with full features configuration (named tsfresh-full) were used with supervised machine

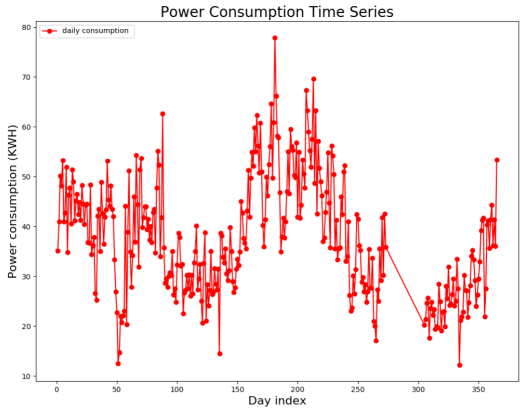


FIGURE 10. Typically power consumption time series.

TABLE 2. Affection comparison of missing data fill methods.

Data set	Fill method	Algorithm	Accuracy
pre-data 1	-1	xgboost	0.9324
pre-data 1	-1	lightGBM	0.9276
pre-data 1	-1	random forest	0.8743
pre-data 2	0	xgboost	0.9328
pre-data 2	0	lightGBM	0.9267
pre-data 2	0	random forest	0.8543
pre-data 3	LIM	xgboost	0.9294
pre-data 3	LIM	lightGBM	0.9186
pre-data 3	LIM	random forest	0.8827

TABLE 3. Performance comparison of HTSF and tsfresh

Method	Sample	Time(s)	Precision	F1
tsfresh-mini	200	4	0.8462	0.7586
tsfresh-mini	500	5	0.6842	0.6667
tsfresh-mini	1000	9	0.6986	0.6667
tsfresh-mini	2000	10	0.7917	0.6786
tsfresh-mini	5000	23	0.7938	0.7454
tsfresh-mini	10000	66	0.8669	0.8355
tsfresh-full	200	1250	1	0.72
tsfresh-full	500	2282	0.6667	0.6027
tsfresh-full	1000	5032	0.7143	0.6667
tsfresh-full	2000	5748	0.7941	0.7297
tsfresh-full	5000	15593	0.7977	0.7457
tsfresh-full	10000	42846	0.8478	0.8125
HTSF	200	32	0.6667	0.6452
HTSF	500	47	0.6667	0.6027
HTSF	1000	118	0.7183	0.6755
HTSF	2000	245	0.8167	0.7
HTSF	5000	579	0.8393	0.7663
HTSF	10000	527	0.8787	0.8257

learning algorithms (xgboost, lightGBM and random forest), as shown in the following table 2. Through this method, it can be seen that the missing-data filling methods may lead to different predicting accuracy, where the necessity of first data extension stage for the three-stage multi-view stacking ensemble has been proved.

One of the main reasons of proposing hierarchical time series feature extracting method is that the existing up-to-date open-source automatic time series relevant features extraction library tsfresh need too much time while extracting features for power consumption anomaly detection. The feature extracting time required for HTSF, tsfresh-mini and tsfresh-full is shown in the following table 3. It's evident that the HTSF need much less time than tsfresh-full. However, the predict performance on the extracted feature sets is better than tsfresh in predicting accuracy.

TABLE 4. Performance comparison of stacking

Model	Accuracy	F1 score
xgboost	0.9224	0.9253
lightGBM	0.9176	0.9245
random forest	0.8743	0.8964
TMSE	0.9278	0.9302

The proposed HTSF based three-stage multi-view stacking ensemble (named TMSE) is compared with meta-models by predicting accuracy and the results shown in the following table. The results illustrate that the proposed HTSF based on three-stage multi-view stacking ensemble method performs better on accuracy and recall.

V. CONCLUSIONS

With the help of industrial internet of things technologies, user power consumption data obtained by distributed sensor and meters were aggregated, and hence the traditional manual detection of electricity stealing was possible to be done by machine learning algorithms. In this paper, electricity stealing user detection solution that has won the competition of Chinese Computer Federation was introduced, which including a systematic time series feature extraction method with a three-stage multi-view stacking ensemble model: the proposed hierarchical time series feature extraction including summary features, shift features, transform features and decompose features, was used to solve the feature engineering issues in anomaly detection of power consumption which outperforms the up-to-date existing time series automatic feature extracting methods, moreover, a three-stage multi-view stacking ensemble model is proposed to help pursue its potential.

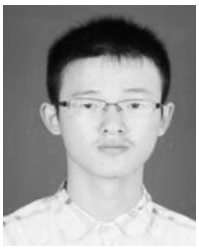
REFERENCES

[1] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 674–686, Mar. 2017.
[2] T. Yijia and G. Hang, "Anomaly detection of power consumption based on waveform feature recognition," in *Proc. 11th Int. Conf. Comput. Sci. Edu. (ICCSE)*, Aug. 2016, pp. 587–591.
[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.
[4] P. J. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.
[5] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 504–509.
[6] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognit. Lett.*, vol. 24, nos. 9–10, pp. 1641–1650, 2003.
[7] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," in *Proc. VLDB Endowment*, vol. 1, no. 2, pp. 1542–1552, 2008.
[8] S. Salvador, P. Chan, and J. Brodie, "Learning states and rules for time series anomaly detection," in *Proc. 17th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2004, pp. 306–311.
[9] T. Yijia and G. Hang, "Anomaly detection of power consumption based on waveform feature recognition," in *Proc. 11th Int. Conf. Comput. Sci. Edu. (ICCSE)*, Aug. 2016, pp. 587–591.

[10] P. Y. Zhou and K. C. C. Chan, "A feature extraction method for multivariate time series classification using temporal patterns," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2015, pp. 409–421.
[11] I. Batal and M. Hauskrecht, "A supervised time series feature extraction technique using DCT and DWT," in *Proc. Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2009, pp. 735–739.
[12] Z. Ouyang, X. Sun, and D. Yue, "Hierarchical time series feature extraction for power consumption anomaly detection," in *Advanced Computational Methods in Energy, Power, Electric Vehicles, and Their Integration*. Singapore: Springer, 2017, pp. 267–275.
[13] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Inf. Fusion*, vol. 40, pp. 45–56, Mar. 2017.
[14] H. Hotelling, "Relations between two sets of variates," *Biometrica*, vol. 28, pp. 321–377, Dec. 1936.
[15] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Conf. Comput. Learn. Theory*, 1998, pp. 92–100.
[16] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.
[17] D. R. Pereira *et al.*, "Social-spider optimization-based support vector machines applied for energy theft detection," *Comput. Electr. Eng.*, vol. 49, pp. 25–38, Jan. 2016.

ZHIYOU OUYANG received the B.E. degree from the School of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, where he is currently pursuing the Ph.D. degree with the School of Automation. His current research interests include machine learning and big data analysis for power systems.

XIAOKUI SUN is currently pursuing the master's degree with the School of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include machine learning and big data analysis for power systems, forecasting issues of renewable generation, and load for power systems.



JINGANG CHEN is currently pursuing the bachelor's degree with the School of Economics, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include machine learning.

DONG YUE (SM'08) received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1995. He is currently a Professor and the Dean with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, and also a Changjiang Professor with the Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. He has published over 100 papers in international journals, domestic journals, and international conferences. His current research interests include analysis and synthesis of networked control systems, multi-agent systems, optimal control of power systems, and Internet of Things. He is currently an Associate Editor of the IEEE Control Systems Society Conference Editorial Board and the *International Journal of Systems Science*.

TENGFEI ZHANG received the bachelor's degree from Henan University, China, in 2002, and the master's and Ph.D. degrees from Shanghai Maritime University, China, in 2004 and 2007, respectively. He is currently a Professor and Supervisor for master's students at the Nanjing University of Posts and Telecommunications. His research interests include intelligent information processing, micro-grid modeling, and control.