**Mathematics and Computer Science**
2018; 3(3): 67-76
http://www.sciencepublishinggroup.com/j/mcs
doi: 10.11648/j.mcs.20180303.11
ISSN: 2575-6036 (Print); ISSN: 2575-6028 (Online)

# Research on Feature Selection in Power User Identification

**Qiu Yanhao[1], Song Xiaoyu[2, *], Sun Xiangyang[2], Zhao Yang[2]**

[1]College of Engineering, Virginia Polytechnic Institute and State University, Virginia, The United States

[2]School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China

**Email address:**
sxy9998@126.com (Song Xiaoyu)

[*]Corresponding author

**To cite this article:**
Qiu Yanhao, Song Xiaoyu, Sun Xiangyang, Zhao Yang. Research on Feature Selection in Power User Identification. *Mathematics and Computer Science.* Vol. 3, No. 3, 2018, pp. 67-76. doi: 10.11648/j.mcs.20180303.11

**Received**: April 18, 2018; **Accepted**: May 7, 2018; **Published**: June 1, 2018

**Abstract:** In the previous study of user identification, most of the researchers improved the recognition algorithm. In this paper, we use large data technology to extract electricity feature from different angles and study the impact of different features on recognition. Firstly, the raw data was cleaned. In order to obtain the key information of power theft user identification, the features of the data set are extracted from three aspects: basic attribute feature, statistical feature under different time scale and similarity feature under different time scale. Then we use feature sets of different combinations to carry out experiments under the KNN model, the random forest (RF) model and the XGBoost model. The experimental results show that the experimental results of the BF+SF+PF feature set in the three classifiers are obviously better than the other two feature sets. Therefore, it is concluded that different features have obvious effects on the recognition results.

**Keywords:** Feature Selection, Power User Identification, KNN, Random Forest, XG Boost

## 1. Introduction

The power industry is the basic industry in the national economy, which is related to the national economy and the people's livelihood, and has the nature of public service. The development of electricity is the basic guarantee for social progress and the improvement of the people's lives. To ensure the development of electricity, it is necessary for power grid enterprises to recover electricity in time，and it is also the inevitable choice to protect the state-owned assets and maintain the market order. However, for a variety of reasons, electricity theft is still widespread, and part of the region is even rampant. The loss of electricity stealing to the power grid enterprises is huge [1]. According to incomp - lete statistics, the economic losses caused by electricity theft in the United States are as high as 4 billion dollars a year. Canada has lost 500 million Canadian dollars a year. Israel loses 40 million dollars a year. And the number in China is up to 20 billion yuan. Southeast Asian countries such as India and Philippines are more serious [2].

With the rapid development of power big data, the use of big data technology to monitor and detect the stealing users of power system has become the trend for the power industry [3]. With the help of big data, the monitoring of electricity users' behavior and identification of electricity stealing behavior can reduce the time and cost of abnormal behavior analysis, improve the recognition rate of abnormal behavior and reduce the operation cost of electric power company.

Many experts at home and abroad have done research on the recognition of electricity stealing users [4-6]. In 2014, Jian Fujun et al quantized the preprocessed data to form feature vectors. Then, the feature vectors were sent to the One-class SVM classifier to identify the users of the electricity stealing [7]. In 2011, Chen et al. used the clustering method of K-means and suffix tree to identify the electricity stealing behavior with using electricity as the feature value [8]. In 2008, Nizar et al. obtained the feature curve of each user by clustering the power consumption curve, and then, according to the deviation degree of the electricity consumption curve and the feature curve, divided the user into two categories: normal and abnormal [9]. However, in the past research, the feature value of the user is relatively simple, and there is no systematic research on the feature value of the user. In the field of data mining, the character value is closely related to the accuracy of the classification. The original data, due to the huge amount of information, is redundant. Extracting feature can not only reduce redundancy, but also make meaningful variables of data become clearer, which is conducive

to later algorithm recognition.

This paper uses large data analysis technology to identify the users of electricity stealing. By comparing experimental results with different feature, it shows that the selection of feature affects the recognition results. Firstly, the original data is preprocessed, including processing the vacancy value and repeated records. The feature are constructed from three aspects: the basic attribute features, the statistical features at different time scales and the similarity features at different time scales. Subsequently, nine groups of experiments are carried out using three different feature sets under the KNN model, the random forest model and the XGBoost model. Finally, the experimental results were analyzed.

## 2. Pearson Correlation Coefficient

The Pearson correlation coefficient, also known as the Pearson product-moment correlation coefficient, is a linear correlation coefficient [10]. The Pearson correlati- on coefficient is used to measure the correlation between the two variables, and the Pearson correlation co- efficient formula is as follows:

$$\begin{aligned}\rho_{X,Y} &= corr(X,Y) \\ &= \frac{\mathrm{cov}(X,Y)}{\delta_X \delta_Y} \\ &= \frac{E[(X-u_X)(Y-u_Y)]}{\delta_X \delta_Y}\end{aligned} \tag{1}$$

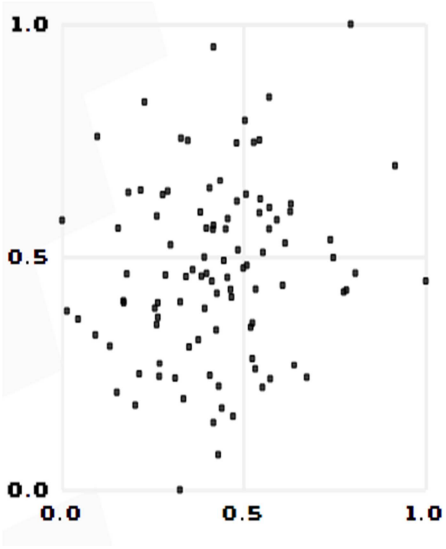


***Figure 1.*** *Data distribution map.*

It is known from the formula that the Pearson correlation coefficient is derived from the standard deviation of the covariance divided by two variables. Although covariance can reflect the correlation degree of two random variables (covariance is greater than 0, it indicates a positive correlation between them. When less than 0, it indicates a negative correlation between them), however, the difference of covariance does not well measure the correlation degree of two random variables. As shown in Figure 1, some data are distributed in a two-dimensional space. It is necessary to analyze the correlation between X axis and Y axis of data points in some problems. If the correlation degree between X and Y is small, but the distribution of data is discrete, it will lead to the bigger covariance difference. It is not reasonable to use this value to measure the correlation degree.

In order to better measure the correlation degree of two random variables, the Pearson correlation coefficient is introduced. It is easy to get that the Pearson correlation coefficient is a value between -1 and 1. When the linear relationship between the two variables is enhanced, the correlation coefficient tends to 1 or -1. When one variable increases and another variable also increases, it shows that they are positive correlation and the correlation coefficient is greater than 0. If one variable increases, but the other is reduced, it indicates that they are negatively correlated, and the correlation coefficient is less than 0. If the correlation coefficient is equal to 0, it shows that there is no linear correlation between them.

## 3. Recognition Model

### *3.1. Data Preprocessing*

The problems of the original data are missing values and repeated records. These problems affect the efficiency of the Classification results, so data preprocessing is the first step of this study. For repeated records, reprocessing is performed according to the main attribute. The treatment of missing values needs careful consideration.

The methods of dealing with the missing value are: ignoring the record, removing the attributes, filling the vacancy manually, using default values, using mean values, inserting values using Lagrange interpolation or Newton interpolation [11]. The Newton interpolation method is used in this paper. Newton interpolation is much simpler. Compared with the Lagrange interpolation, it not only over- comes the shortcoming that the whole calculation work must be restarted when adding a node, but also saves the multiplication and division operation times.

### *3.2. Feature Extraction*

Feature extraction refers to the extraction of key features from the original data as needed. In order to extract more useful information, it is necessary to build new attributes on the basis of existing attributes. In this paper, feature extraction is carried out from the following aspects:

(1) Basic attribute features (BF)

The statistical features of each user ID are recorded, including maximum, minim- um, mean, variance, median and number of records and so on. The number of records is closely related to the statistical features. These are the basic features of electricity user.

(2) Statistical features at different time scales (SF)

One is electricity consumption of the user in every 3 days, per month, 3 months and half year. The other is record number of the user in every 3 days, per month, 3 months and half year. The

features of different time scales provide important information for the detection of different types of abnormal user.

(3) Similarity features at different time scales (PF)

The Pearson Correlation Coefficient is used to measure the correlation between two variables. The Pearson correlation coefficient of power consumption, power starting degree and power termination degree are calculated during per 4 weeks and 5 weeks.

# 4. Results and Evaluation

### *4.1. Data Set*

The sample data comes from the State Grid Gansu Electric Power Company, which contains 6 million 300 thousand electricity consumption records of 9956 power customers. 1394 of the customers have been identified as stealing users through offline investigation, and the rest are normal users. Each user's data includes the user's daily electricity consumption, the day and yesterday's electricity consumption.

### *4.2. Evaluation Method*

#### *4.2.1. Confusion Matrix*

The confusion matrix is the analysis table that summarizes the classification model prediction results in machine learning [12-13]. In the form of matrix, the records of the data set are collected in accordance with the two criteria, which are the real category and the categories predicted by the classification. The confusion matrix shown in Table 1 shows all possible classification results of the classifier. Each column represents the predicted value, and each row represents the actual category. User identification of electricity stealing is essentially a binary classifier classification problem, and all users are divided into two categories: normal users and stealing users. In this paper, positive and negative correspond to stealing users and normal users respectively.

***Table 1.*** *The confusion matrix form.*

| | | **actual value** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| Predicted value | Positive | True positive (TP) | False positive (FP) |
| | Negative | False negative (FN) | True negative (TN) |

FP is the first type of error, and FN is the second type. On the basis of confusion matrix, the evaluation indexes of classifiers can be derived: accuracy (ACC) and the true positive rate (TPR) [14]. Accuracy describes the classification accuracy of the classifier. The true positive rate is also called sensitivity. It describes the sensitivity of classifiers.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \qquad (2)$$

$$TPR = \frac{TP}{TP + FN}$$

#### *4.2.2. ROC Curve and AUC*

The ROC curve (receiver operating feature curve, the feature curve of the subject) is a graphical method of displaying the compromise between the true rate (TPR) and the false positive rate (FPR) of the classifier [15]. In a two classification model, for the continuous result, it is assumed that a threshold has been determined. For example, 0.6, the instance that is greater than this value is classified as a positive class, and the instance less than that is assigned to the negative class. If the threshold is reduced to 0.5, more classes can be identified. This is to improve the ratio of positive cases to all positive cases, that is TPR, but at the same time, more negative examples are also taken as a positive example, that is, the increase of FPR. In order to visualize this change, the ROC is introduced here, and the ROC curve can be used to accurately evaluate a classifier. The curve closed point (0, 1) shows the best classification effect [16].

### *4.3. Experimental Results*

The software used in this experiment is R on the computer with Intel Corei5-4210, 2.4 Ghz, 8 G, win10x64. Using three different combinations of features, experiments were carried out under KNN classifier, random forest classifier and XGBOOST classifier.

There are nine groups of experiments in this paper, the first three groups are carried out using three different feature sets under the KNN model, and the middle three groups are carried out using three different feature sets under the random forest model. The last three groups are carried out using three different feature sets under the XGBoost model.

The experiments are conducted 5 times and carried out under different recognition models. KNN classifier takes K=50. In the BP neural network, the hidden layer unit is 8. The training algorithm used Quick Prop. The algorithm parameters are 0.1, 2, 0.0001, 0.1. The maximum number of iterations is 1000. The XGBoost parameters are divided into three kinds: general parameters, booster parameter and learning target parameter [17-18]. In this experiment, the booster parameter shrinkage step size (ETA) is selected 0.01 to prevent overfitting, and the maximum iteration number (nrounds) is selected 1500. The minimum sample weight of child nodes (min child weight) is selected 10. The ratio of feature sampling (colsample by tree) is 0.8. In the learning target parameters, the objective function selects binary logistic regression (binary logistic), and the evaluating indicator is the average accuracy (map). The rest parameters keep the default value. The results are as follows.

The first group of experiments used features of three different combinations to test under the KNN model. the experimental results are as follows:

***Table 2.*** *Results of five experiments under BF--KNN model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF | KNN | 1 | 63.67% | 40.29% |
| | | 2 | 65.47% | 37.77% |
| | | 3 | 63.13% | 35.25% |
| | | 4 | 65.29% | 38.49% |
| | | 5 | 64.39% | 35.97% |
| | | Average | 64.39% | 37.55% |

***Table 3.*** *Results of five experiments under BF+SF--KNN model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF | KNN | 1 | 73.02% | 53.60% |
| | | 2 | 68.35% | 46.40% |
| | | 3 | 69.06% | 48.92% |
| | | 4 | 69.06% | 50.36% |
| | | 5 | 68.17% | 47.84% |
| | | Average | 69.53% | 49.42% |

***Table 4.*** *Results of five experiments under BF+SF+PF--KNN model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF+PF | KNN | 1 | 73.92% | 54.68% |
| | | 2 | 73.02% | 52.52% |
| | | 3 | 74.82% | 55.76% |
| | | 4 | 74.46% | 54.32% |
| | | 5 | 74.82% | 54.68% |
| | | Average | 74.21% | 54.39% |

From the above table, it can be seen that the average ACC and TPR of the KNN model are 64.39% and 37.55% respectively, when only BF is used. The average ACC and TPR of the KNN model are 69.53% and 49.42% respectively, when BF and SF are used. The average ACC and TPR of the KNN model are 74.21% and 54.39% respectively, when BF, SF and PF are used. Through the above results, it can be seen that, under the KNN model, the difference of ACC and TPR is obvious under different features. Under BF+SF+PF, the recogni- tion results are best, followed by BF+SF, and BF is the worst.

Figure 2-4 is the ROC curve of the three experiments. From the ROC curve, it can be seen that under different characteristics, the degree of curve approaching 1 is also different. Among them, BF+SF+PF is the best and BF is the worst.

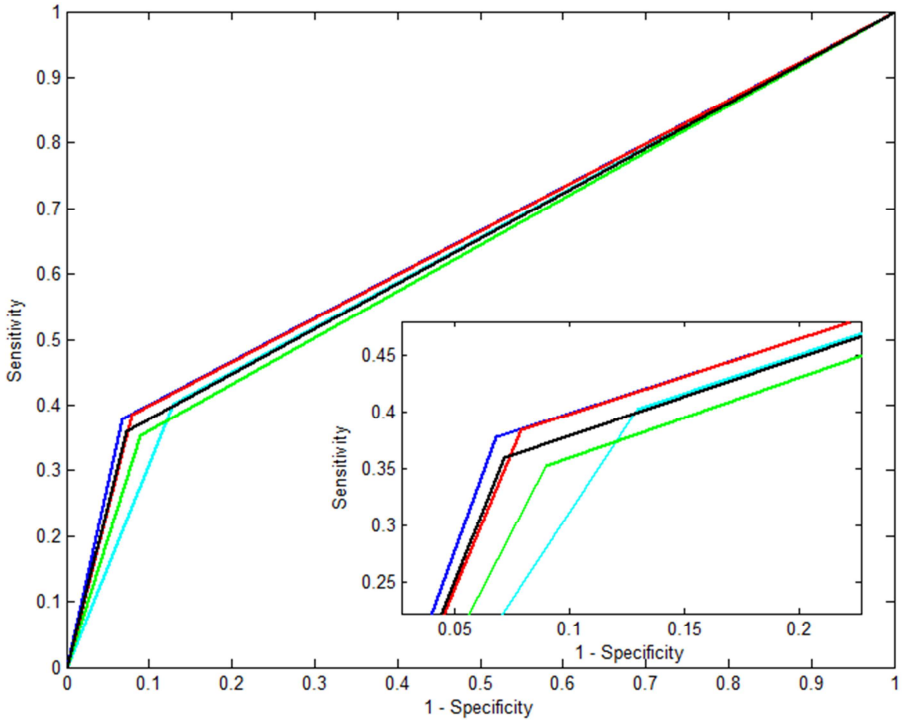


***Figure 2.*** *ROC curve of BF--KNN model.*

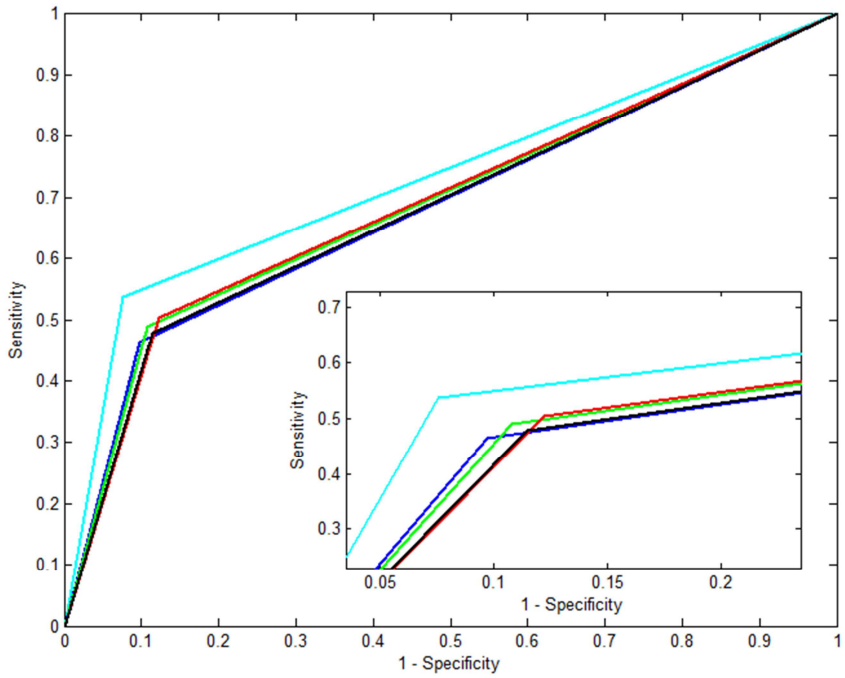


***Figure 3.*** *ROC curve of BF+SF--KNN model.*

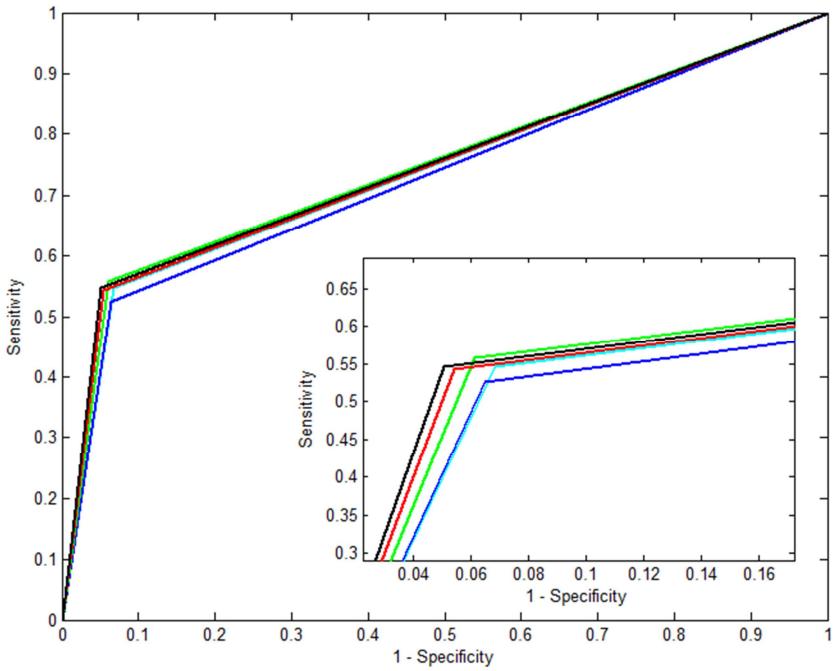


***Figure 4.*** *ROC curve of BF+SF+PF--KNN model.*

The second group of experiments used features of three different combinations to test under the random forest model. The experimental results are as follows.

***Table 5.*** *Results of five experiments under BF--RF model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF | RF | 1 | 71.22% | 51.08% |
| | | 2 | 70.68% | 52.52% |
| | | 3 | 69.42% | 45.68% |
| | | 4 | 69.06% | 45.32% |
| | | 5 | 67.81% | 46.40% |
| | | Average | 69.64% | 48.20% |

***Table 6.*** *Results of five experiments under BF+SF--RF model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF | RF | 1 | 72.30% | 52.88% |
| | | 2 | 72.12% | 52.16% |
| | | 3 | 72.30% | 51.08% |
| | | 4 | 71.94% | 50.72% |
| | | 5 | 73.56% | 55.04% |
| | | Average | 72.45% | 52.37% |

***Table 7.*** *Results of five experiments under BF+SF+PF--RF model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF+PF | RF | 1 | 78.96% | 75.90% |
| | | 2 | 79.14% | 74.10% |
| | | 3 | 78.60% | 76.26% |
| | | 4 | 80.76% | 74.82% |
| | | 5 | 80.58% | 78.06% |
| | | Average | 79.60% | 75.83% |

From the above table, it can be seen that the average ACC and TPR of the random forest model are 69.64% and 48.20% respectively, when BF is used. The average ACC and TPR of the random forest model are 72.45% and 52.37% respectively, when BF and SF are used. The average ACC and TPR of the random forest model are 79.60% and 75.83% respectively, when BF, SF and PF are used. when BF, SF and PF are used. Through the above results, it can be seen that, under the random forest model, the difference of ACC and TPR is obvious under different features. Under BF+SF+PF, the recognition results are best, followed by BF+SF, and BF is the worst.

Figure 5-7 is the ROC curve of the three experiments. From the ROC curve, it can be seen that under different characteristics, the degree of curve approaching 1 is also different. Among them, BF+SF+PF is the best and BF is the worst.

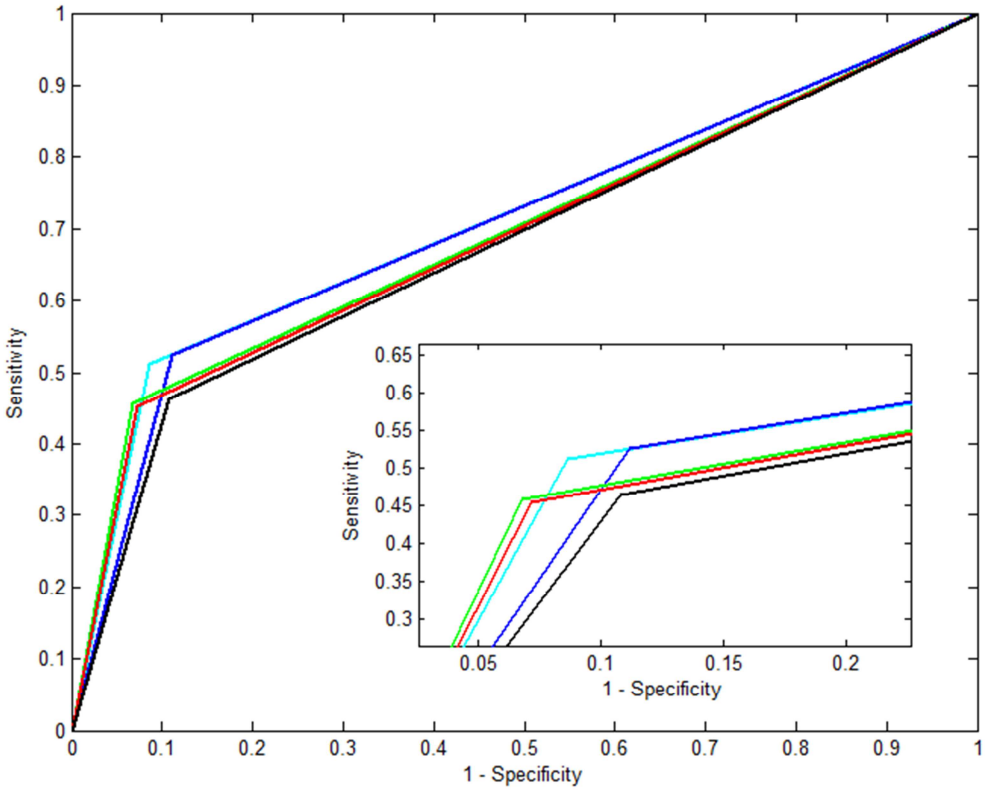


***Figure 5.*** *ROC curve of BF--RF model.*

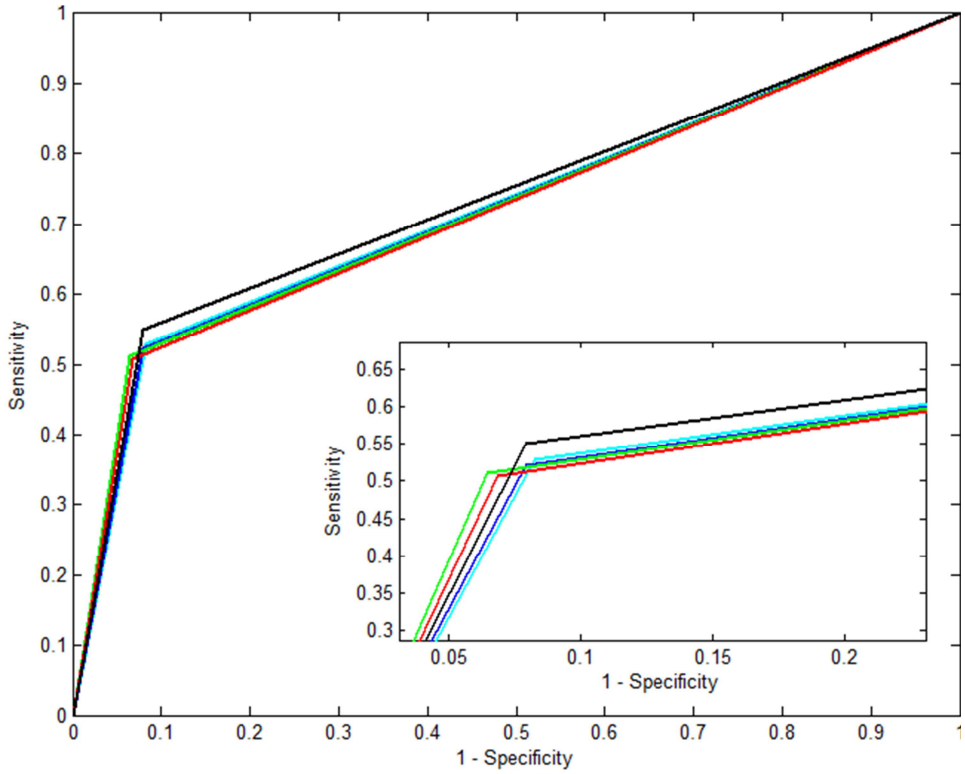


***Figure 6.*** *ROC curve of BF+SF--RF model.*

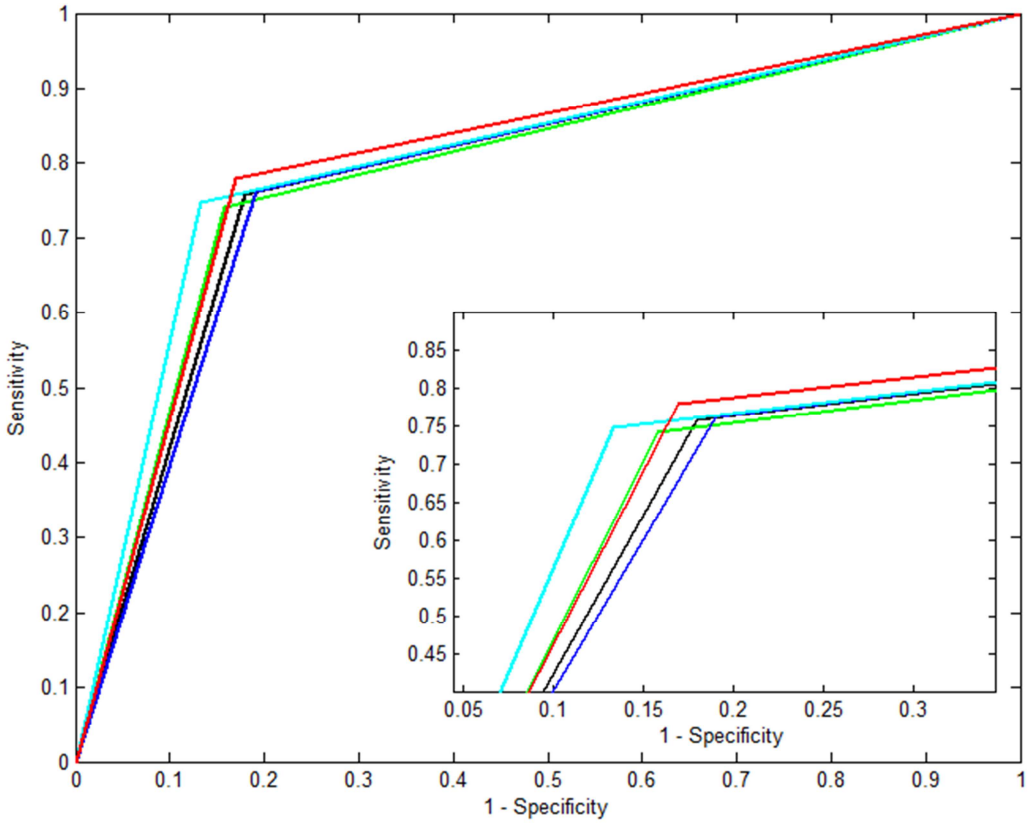


***Figure 7.*** *ROC curve of BF+SF+PF--RF model.*

The third group of experiments used features of three different combinations to test under the XGBoost model. The experimental results are as follows:

***Table 8.*** *Results of five experiments under BF--XGBoost model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF | XGBoost | 1 | 78.78% | 78.42% |
| | | 2 | 76.44% | 74.10% |
| | | 3 | 77.34% | 75.90% |
| | | 4 | 78.42% | 75.18% |
| | | 5 | 76.26% | 77.70% |
| | | Average | 77.45% | 76.26% |

***Table 9.*** *Results of five experiments under BF+SF--XGBoost model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF | XGBoost | 1 | 79.68% | 76.98% |
| | | 2 | 81.12% | 77.34% |
| | | 3 | 81.12% | 75.18% |
| | | 4 | 81.29% | 75.54% |
| | | 5 | 80.94% | 77.34% |
| | | Average | 80.83% | 76.47% |

***Table 10.*** *Results of five experiments under BF+SF+PF--XGBoost model.*

| Features | Classifier | No | ACC | TPR |
|---|---|---|---|---|
| BF+SF+PF | XGBoost | 1 | 81.83% | 75.18% |
| | | 2 | 80.76% | 78.78% |
| | | 3 | 81.83% | 78.78% |
| | | 4 | 81.29% | 77.70% |
| | | 5 | 81.83% | 78.42% |
| | | Average | 81.51% | 77.77% |

From the above table, it can be seen that the average ACC and TPR of the XGBoost model are 77.45% and 76.26% respectively, when only BF is used. The average ACC and TPR of the XGBoost model are 80.83% and 76.47% respectively, when BF and SF are used. The average ACC and TPR of the XGBoost model are 81.51% and 77.77% respectively, when BF, SF and PF are used. when BF, SF and PF are used. Through the above results, it can be seen that, under the XGBoost model, the difference of ACC and TPR is obvious under different features. Under BF+SF+PF, the recognition results are best, followed by BF+SF, and BF is the worst.

Figure 8-10 is the ROC curve of the three experiments. From the ROC curve, it can be seen that under different characteristics, the degree of curve approaching 1 is also different. Among them, BF+SF+PF is the best and BF is the worst.

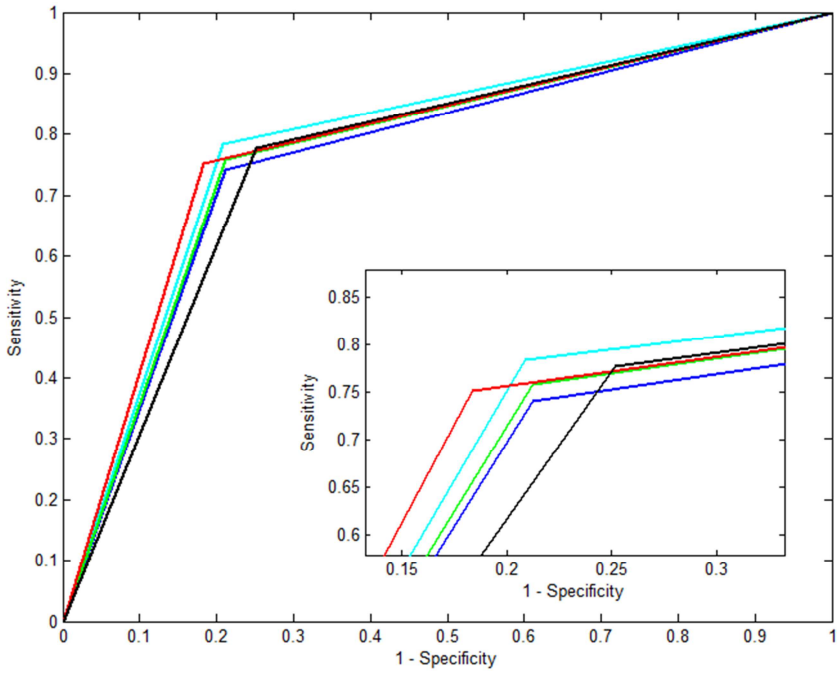


***Figure 8.*** *ROC curve of BF--XGBoost model.*

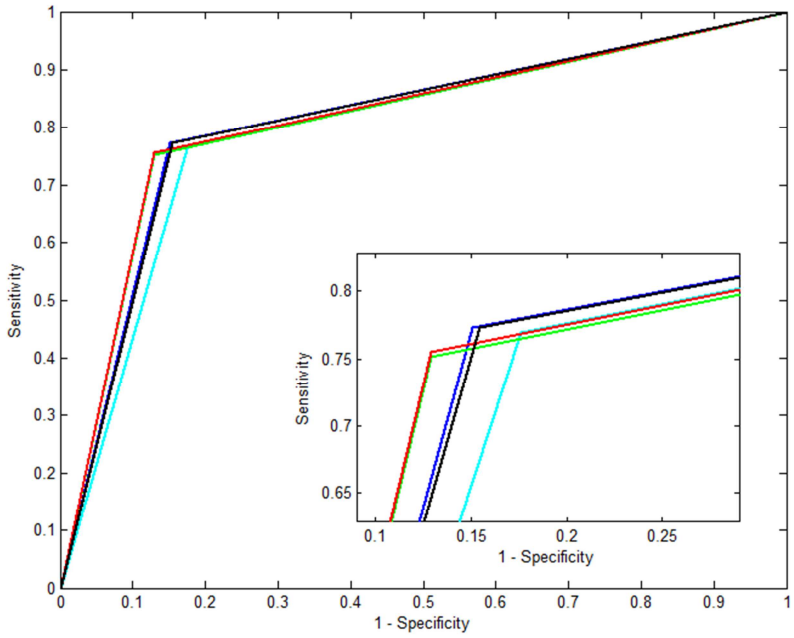


***Figure 9.*** *ROC curve of BF+SF--XGBoost model.*

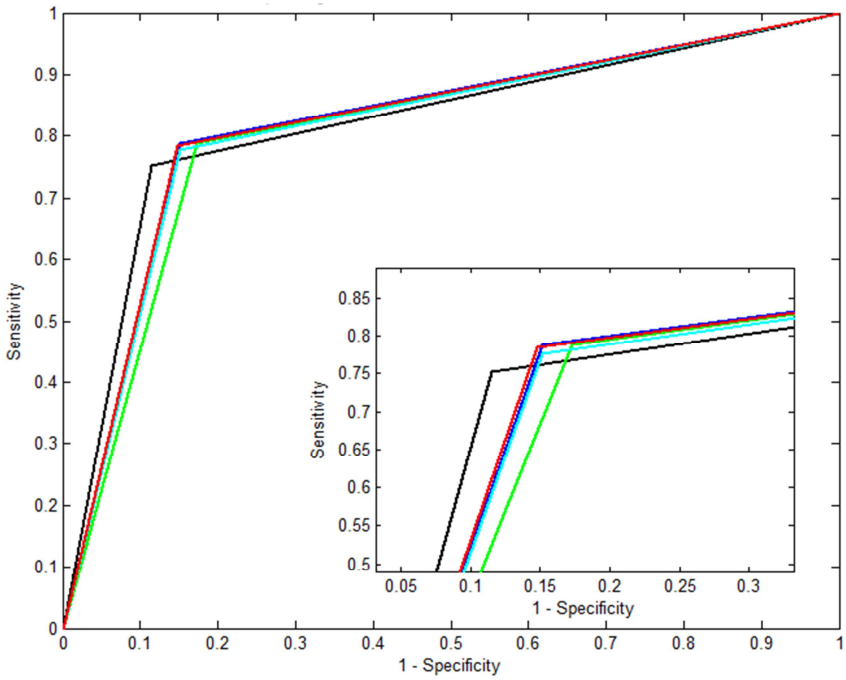


***Figure 10.*** *ROC curve of BF+SF+PF--XGBoost model.*

In order to compare, tables 11 and 12 are obtained by summarizing the results. The values in the table are the average values of each set of experimental results.

***Table 11.*** *ACC value of evaluation indicators.*

| ACC | KNN | RF | XGBoost |
|---|---|---|---|
| BF | 64.39% | 69.64% | 77.45% |
| BF+SF | 69.53% | 72.45% | 80.83% |
| BF+SF+PF | 74.21% | 79.60% | 81.51% |

It can be seen from the table 11 that ACC of BF+SF+PF feature set increases by 9.82% and 4.68% compared with BF feature set and BF+SF feature set under KNN classifier. ACC of BF+SF+PF feature set increases by 9.96% and 7.15% compared with BF feature set and BF+SF feature set under random forest classifier. ACC of BF+SF+PF feature set increases by 4.06% and 0.68% compared with BF feature set and BF+SF feature set under XGBoost classifier. BF+SF+PF.

***Table 12.*** *TPR value of evaluation indicators.*

| TPR | KNN | RF | XGBoost |
|---|---|---|---|
| BF | 37.55% | 48.20% | 76.26% |
| BF+SF | 49.42% | 52.37% | 76.47% |
| BF+SF+PF | 54.39% | 75.83% | 77.77% |

It can be seen from the above table 12 that TPR of BF+SF+PF feature set increases by 16.84% and 4.97% compared with BF feature set and BF+SF feature set under KNN classifier. TPR of BF+SF+PF feature set increases by 27.63% and 23.46% compared with BF feature set and BF+SF feature set under random forest classifier. TPR of BF+SF+PF feature set increases by 1.51% and 1.3% compared with BF feature set and BF+SF feature set under XGBoost classifier. BF+ SF+ PF.

Therefore, through the above results, it can be concluded that the selection of features has a great influence on the recognition results.

# 5. Conclusion

With the continuous improvement of information degree in power system and the rapid growth of electricity consumption data, it is important to study the analysis technology suitable for power big data, which is of great significance for the innovation of power business mode and the development of smart grid.

In this paper, the features of three different combinations are constructed under the KNN model, the random forest model and the XGBoost model, nine groups of experiments are carried out to study the effect of feature selection on the identification of electricity stealing user. From the experimental results, it can be seen that the characteristics of different combinations have obvious impact on the experimental results, the more features that can characterize the behavior of users, have better results. From another point of view, it can be found that different classifiers have influence on the recognition results. In further, more effective feature will be obtained to improve the accuracy of recognition.

# Acknowledgements

This work is supported by National Natural Science Foundation of China (61262044).

# References

[1] Song Y, Zhou G, Zhu Y. Present Status and Challenges of Big Data Processing in Smart Grid. Power System Technology, 2013, 37(4):927-935.

[2] Tan Z. Design and implementation of online abnormal electricity utilization and risk monitoring system based on electricity behavior analysis. South China University of Technology, 2015.

[3] Chen W, Chen Y, Qiu L, et al. Analysis of anti-stealing electric power based on big data technology. Journal of Electronic Measurement & Instrumentation, 2016.

[4] Zhuang C, Zhang B, Jun H U, et al. Anomaly Detection for Power Consumption Patterns Based on Unsupervised Learning. Proceedings of the Csee, 2016.

[5] Zhou L, Zhao L, Gao W. Application of Sparse Coding in Detection for Abnormal Electricity Consumption Behaviors. Power System Technology, 2015

[6] Monedero I, Biscarri F, León C, et al. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. International Journal of Electrical Power & Energy Systems, 2012, 34(1): 90-98.

[7] Jian F J, Cao M, Wang L, et al. SVM Based Energy Consumption Abnormality Detection in AMI System. Electrical Measurement & Instrumentation, 2014.

[8] Chen C, Cook D J. Energy Outlier Detection in Smart Environments// Artificial Intelligence and Smarter Living: the Conquest of Complexity, Papers From the 2011 AAAI Workshop, San Francisco, California, Usa, August. 2011.

[9] Nizar A H, Dong Z Y, Wang Y. Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 2008, 23(3): 946-955.

[10] Geng Y J, Zhang J Y, Yuan X G. A feature relevance measure based on sparse representation coefficient. Pa- ttern Recognition & Artificial Intelli- gence, 2013, 26(1):106-113.

[11] Zhang Y, Shang C. Combining Newton interpolation and deep learning for image classification. Electronics Letters, 2015, 51(1):40-42.

[12] Kong Y H, Jing M L. Research of the Classification Method Based on Confu- sion Matrixes and Ensemble Learning. Computer Engineering & Science, 2012, 34(6):111-117.

[13] Song Y F, Wang X D, Lei L. Evaluating evidence reliability based on confusion matrix. XI Tong Gong Cheng Yu Dian Zi Ji Shu/systems Engineering & Elec- tronics, 2015, 37(4):974-978.

[14] Huang Y A, You Z H, Gao X, et al. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. Biomed Research International, 2015, 2015:902198.

[15] Song H L, He J, Huang P X, et al. Application of parametric method and non-parametric method in estimation of area under ROC curve. Academic Journal of Second Military Medical University, 2006, 27(7):726-728.

[16] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8):861-874.

[17] Chen T, Guestrin C. XGBoost:A Scalable Tree Boosting System// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.

[18] Zhang L, Zhan C. Machine Learning in Rock Facies Classification: An Applic- ation of XGBoost// International Geophysical Conference, Qingdao, China, 17-20 April. 2017: 1371-1374.