Abnormal user identification based on XGBoost algorithm

SONG Xiao-yu，SUN Xiang-yang，ZHAO Yang

(*School of Electronic and Information Engineering，Lanzhou Jiaotong University，Lanzhou 730070，China*)

Abstract：The eXtreme gradient boosting (XGBoost) algorithm is used to identify abnormal users. Firstly, the raw data were cleaned. Then user power characteristics were extracted from different aspects. Finally, the XGBoost classifier was used to identify the abnormal users respectively in the balanced sample set and the unbalanced sample set. In contrast，under the same characteristics，the k-nearest neighbor (KNN) classifier，back-propagation (BP) neural network classifier and random forest classifier were used to identify the abnormal users in the two samples. The experimental results show that the XGBoost classifier has higher recognition rate and faster running speed. Especially in the imbalanced data sets, the performance improvement is obvious.

Key words：user identification；electricity characteristics；eXtreme gradient boosting (XGBoost)；random forest

0　Introduction

The identification of abnormal users is the most difficult problem in the traditional power industry. With the rapid development of economy，social production and people's daily electricity consumption have increased year by year. Due to illegal use of electricity by some people，the abnormal phenomenon of power supply is becoming more and more serious[1]，which not only brings the economic loss to power supply enterprises，but also affects normal power supply and stability.

In the electric power industry，the strong smart grid is rapidly developing with an unprecedented breadth and depth. The arrival of big data and cloud computing era injects new vitality into the development of traditional power industry[2]. The purpose of this paper is to use the big data algorithm to intelligently identify the abnormal electricity customers. With the help of big data technology，the monitoring of abnormal behavior and the recognition of abnormal power users can reduce the cost of abnormal behavior analysis and improve the recognition rate of abnormal behavior.

For the detection of electricity users，many domestic and foreign experts use data mining technology[3-4]. In 2016，Zhang，et al. extracted a variety of characteristic quantities that could characterize the user power consumption model from the electricity consumption data. The power consumption sequence of all users was mapped to two-dimensional plane by dimension reduction first，and then the local outlier factor algorithm was used to calculate the outlier degree of each user，finally the abnormal users were identified[5]. In 2015，Zhou，et al proposed a sparse coding model method to mine the users' original electricity data. Combined with the method of dictionary learning，user data were represented as a linear combination of features. By using the frequency of each feature to judge the abnormal value，the abnormal behavior was distinguished[6]. In 2012，Monedero，et al. used the Pearson correlation coefficient to detect the typical loss of electricity characterized by the sudden drop in electricity consumption. For other types of power loss，the Bayesian network and decision tree were used[7]. However，in real life，the abnormal users generally are in a minority，and abnormal users determined by artificial field must be in a minority，which led to the imbalance of positive and negative samples. In previous supervised learning research，most of researchers did not make research on imbalanced sample set and optimize the speed of model operation. As a result，power industry processing data records often reach tens of millions，

even hundreds of millions，which makes most identification models run longer in practical applications.

In this paper，large data analysis algorithm are used to identify abnormal power users. Firstly，the raw data are pre-processed，including vacancy values and duplicate records. Then the feature engineering is constructed from three aspects：statistical features，different time scale features and correlation coefficient features. After that，the eXtreme gradient boosting (XGBoost) classifier is used to identify the features. Finally，the classification results are compared with that of k-nearest neighbor (KNN) classifier，back-propagation (BP) neural network classifier and random forest classifier.

1　XGBoost algorithm

The XGBoost is an extension of the gradient boosting decision tree algorithm[8-9]. Boosting classifier belongs to ensemble learning model，and its basic idea is to combine hundreds of tree models with lower classification accuracy. The model is iterated，and each iteration generates a new tree. How to generate a reasonable tree at each step is the core of the boosting classifier. Gradient boosting machine algorithm uses the idea of gradient descent when generating each tree. All the trees generated by the above steps are based on the direction of minimizing the given objective function. Under the reasonable parameter setting，a certain number of trees can reach the expected accuracy. XGBoost algorithm is the implementation of gradient boosting machine，which can automatically use CPU multi-threaded parallel[10]. The specific theory is as follows.

Decision tree is widely used for classification model because of its intuitive representation and reliable basis. However，the performance of single decision tree is limited，so the decision tree ensemble is usually used to improve the performance. Given data set $D = \{(x_i, y_i)\}(|D| = n, x_i \in \mathbf{R}^m, y_i \in \mathbf{R})$. The integrated model of trees is expressed as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F,$$

$$F = \{f(x) = \omega_{q(x)}\}(q : \mathbf{R}^m \rightarrow T, \omega \in \mathbf{R}^T), \quad (1)$$

where k is the number of models in the integrated model，F is a set space of regression trees，x_i represents the eigenvectors of the first i data points，q represents the index of the leaf，T represents the number of leaves on a tree，every $f_k(\cdot)$ corresponds to an independent tree structure $q(\cdot)$ and the weight of the leaf ω.

The objective function consists of two parts，that is

$$O(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad (2)$$

where the first part $l(\cdot)$ is the training error between the predicted value $\hat{y}_i$ and real value y_i of the target，and the second part $\Omega(\cdot)$ is the sum of the complexity of each tree as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel \omega \parallel^2, \quad (3)$$

where λ and γ are coefficients. The minimum value of the objective function is its optimal value.

In Eq. (2)，the objective function of the integrated decision tree model is trained by boosting method，that is to say，each time the model is retained，and a new function is added to the model as

$$\hat{y}_i^{(0)} = 0,$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i),$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i),$$

$$\vdots$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (4)$$

where $\hat{y}_i^{(t)}$ is the predicted value of the sample i in the round t；$\hat{y}_i^{(t)}$ retains the predicted value of the model $\hat{y}_i^{(t-1)}$ in the round $t-1$ and then adds a new function $f_t(x_i)$. The choice of adding a new function in each round is to minimize the objective function as much as possible. Thus，the objective function turns to

$$O^{(t)} = L^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) =$$

$$\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C. \quad (5)$$

The objective function can be optimized through $f_t(\cdot)$，When the error function $l(\cdot)$ is squared error，the objective function can be written as

$$L^{(t)} = \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i)f_t(x_i) + f_t(x_i)^2] +$$

$$\Omega(f_t) + C. \quad (6)$$

For other error functions in addition to squared errors，Taylor expansion is used to approximately define the objective function as

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), \quad h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)}), \quad (7)$$

$$\widetilde{L}^{(t)} \cong \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + C. \quad (8)$$

After the constant term is removed, a relatively uniform objective function is obtained as

$$\widetilde{L}^{(t)} \cong \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t). \quad (9)$$

Based on the above theory, the XGBoost algorithm can be described as follows:

1) Initialize a new decision tree.

2) Calculate the first derivative and the second derivative of each training sample point.

3) Calculate the residual of samples in function space.

4) Turning towards gradient direction of residual reduction, a new decision tree is generated by greedy strategy.

5) Add the newly generated decision tree to the model and update the model.

6) Back to step 2, recursively execute until the objective function satisfies the condition.

2 Detection model

2.1 Data preprocessing

The problems of the original data are missing values and repeated records. These problems affect the efficiency of the classification results, therefore 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 include ignoring the record, removing the attributes, filling the vacancy manually, using default values, using mean values and inserting values by means of Lagrange interpolation or Newton interpolation. The Newton interpolation method is used in this paper since it is much simpler compared with Lagrange interpolation. It not only overcomes the shortcoming that the whole calculation work must be restarted when adding a node, but also saves the multiplication and division operation times.

2.2 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 characteristics

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

2) Characteristics on different time scales

One is electricity consumption of the user in every three days, one month, three months and half a year. The other is the number of the user records in every three days, one month, three months and half a year. The characteristics of different time scales provide important information for the detection of different types of abnormal users.

3) Similarity features on different time scales

The Pearson correlation coefficient is used to measure the correlation between two variables[11]. The Pearson correlation coefficients of power consumption, power starting degree and power termination degree are calculated during per four weeks and five weeks.

3 Results and evaluation

3.1 Data set

The sample data come from the State Grid Gansu Electric Power Company, which contain 6.3 million electricity consumption records from 9 956 users customers, 1 394 of which have been identified as abnormal users through offline investigation, and the rest are normal users. The data from each user include his daily electricity consumption as well as electricity consumption of the day and yesterday. The proportion of positive and negative sample data is about 1 : 6, which makes the feature extraction and recognition of abnormal users become much difficult.

3.2 Evaluation method

3.2.1 Confusion matrix

Essentially, abnormal user identification is a two-element classification problem and it divides all users into two categories: normal users and abnormal users. For the two-element classification problem, confusion matrix is a basic tool to evaluate the

reliability of classifier. The confusion matrix in Table 1 shows all possible classification results of the classifier. The rows correspond to the categories to which the object actually belongs, and the lists show the categories of the prediction. In this paper, "positive" and "negative" correspond to abnormal users and normal users respectively.

Table 1 Confusion matrix form

		Actual value	
		Positive	Negative
Predicted value	Positive	True positive (V_{TP})	False positive (V_{FP})
	Negative	False negative (V_{FN})	True negative (V_{TN})

V_{FP} refers to the first type error, and V_{FN} refers to the second type error. On the basis of confusion matrix, the evaluation indices of classifiers can be derived as follows: accuracy, true positive rate and specificity. The accuracy (V_{ACC}) describes the classification accuracy of the classifier. The true positive rate (V_{TPR}) also called sensitivity describes the sensitivity of the classifier. The specificity (V_{SPC}) describes the correct prediction of the negative in all negative samples.

$$V_{ACC} = \frac{V_{TP} + V_{TN}}{V_{TP} + V_{FP} + V_{TN} + V_{FN}},$$

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

$$V_{SPC} = \frac{V_{TN}}{V_{TN} + V_{FP}}. \tag{10}$$

3.2.2 Receiver operating characteristic curve and area under curve

Receiver operating characteristic (ROC) curve describes the relative relationship between the growth rates of the two indices in the confusion matrix: false positive rate (V_{FPR}) and V_{TPR}[12]. For the continuous value of the output two-element classification model, the samples larger than the threshold are classified as positive class, and the samples less than the threshold are classified as negative class. Reducing the threshold can recognize more positive classes and V_{TPR} will increase, but at the same time, more negative samples will be classified as positive class wrongly and V_{FPR} will increase. The ROC curve can represent the process of change. The curve closed point (0, 1) shows the best classification effect. Area under curve (AUC) can indicate the quality of classifier. The much larger AUC represents the better performance. If AUC is equal to 1, it means an ideal classifier.

3.3　Experimental results

The software used in this experiment is R on the computer with Intel Corei5-4210, 2.4 GHz, 8 GB, win10x64. The experiments were carried out in two data sets using KNN classifier, random forest classifier and XGBoost classifier, respectively. A group of experiments using a balanced sample set was carried out, and the other group was carried out on a sample set with imbalanced sample set. In order to ensure the fairness of the experiments, the eigenvalues of each model were the same.

3.3.1　Balanced sample set

The balanced sample set was composed of 1 394 positive samples and 1 394 negative samples. The former samples were the existing abnormal users, and the latter samples were the samples randomly selected from 8 562 normal users. The balanced sample set was divided into five parts, four of which were training sets, one of which was a test set. The experiments were conducted five times under different recognition models. KNN classifier taked K = 50. In the BP neural network, the number of hidden layer unit was 8. The training algorithm used QuickProp. The algorithm parameters were 0.1, 2, 0.000 1 and 0.1. The maximum number of iterations was 1 000. The XGBoost parameters were divided into three kinds: general parameters, booster parameters and learning target parameters. In this experiment, 0.01 was selected as the booster parameter shrinkage step size (ETA) to prevent overfitting, and the maximum iteration number (nrounds) was 1 500. The minimum sample weight of child nodes (min_child_weight) was 10. The ratio of feature sampling (colsample_bytree) was 0.8. In the learning target parameters, the objective function selected binary logistic regression (binary: logistic), and the evaluating indicator was the average accuracy (map). The rest parameters kept the default values. The results are shown in Table 2 and the values are the average of the five experiments.

Table 2 Results of five experiments under different models

	$V_{ACC}(\%)$	$V_{TPR}(\%)$	$V_{SPC}(\%)$	AUC(%)	Time(s)
KNN	69.53	49.42	89.64	69.54	234
BP	73.24	70.00	76.47	73.24	232
RF	79.60	75.83	83.38	79.61	249
XGBoost	81.51	77.77	85.25	81.52	212

From the above table, we can see that the average V_{ACC}, V_{TPR}, V_{SPC} and the time of the KNN model are 69.53%，49.42%，89.64% and 234 s, respectively. The average V_{ACC}, V_{TPR}, V_{SPC} and the time of BP neural network model are 73.24%，70%，76.47% and 232 s, respectively. The mean V_{ACC}, V_{TPR}, V_{SPC} and the time spent on random forest model are 79.60%，75.83%，83.38% and 249 s, respectively. Finally, the average V_{ACC}, V_{TPR}, V_{SPC} and the time spent on the XGBoost model are 81.51%，77.77%，85.25% and 212 s, respectively. It can be seen that the XGBoost classifier are better than the other three classifiers. Figs. 1-4 are ROC curve graphs of KNN model、BP neural network、random forest model and XGBoost model, respectively. The AUC of these models are 69.54%，73.24%，79.61%，81.52%, respectively. The higher AUC value represents a better classification effect.
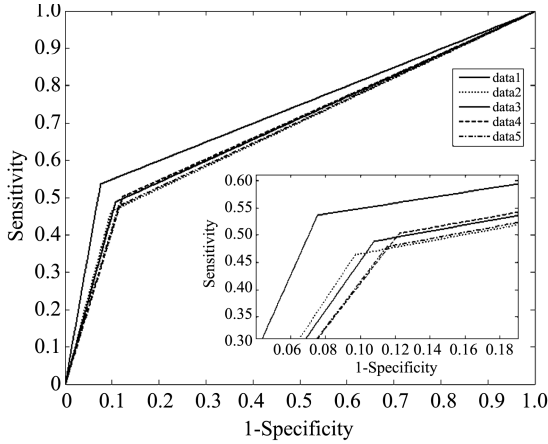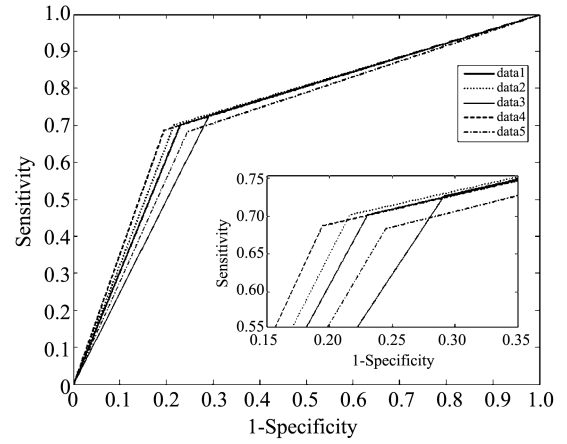

Fig. 1   ROC curve of KNN model

Fig. 2   ROC curve of BP neural network model
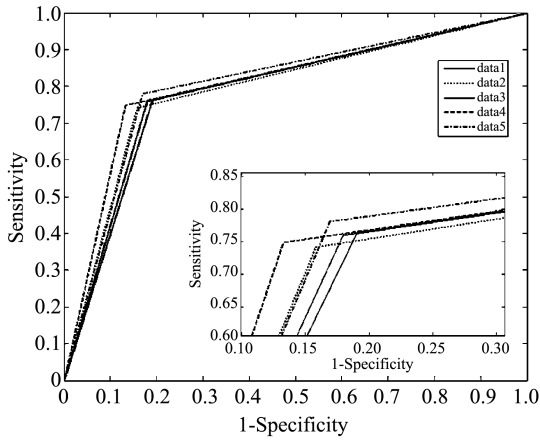

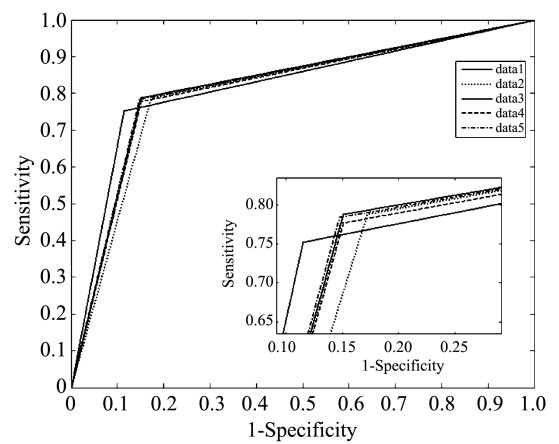**Fig. 3   ROC curve of random forest model**

Fig. 4   ROC curve of XGBoost model

3.3.2   Imbalanced sample set

The imbalanced sample set consisted of 1 394 positive samples and 8 562 negative samples. The former samples were the existing abnormal users, and the latter samples were all normal users. The imbalanced sample set was divided into five parts, four of which were training sets, one of which was a test set. The experiments were conducted five times under different recognition models. KNN classifier taked $K = 100$. In the BP neural network, the number of hidden layer units was 5. The training algorithm used QuickProp. The algorithm parameters were 0.1，2，0.000 1 and 0.1. The maximum number of iterations was 100. The XGBoost parameters were divided into three kinds: general parameters, booster parameters and learning target parameters. In this experiment, the booster parameter shrinkage step size（ETA）was 0.01 to prevent overfitting, and the maximum iteration number（nrounds）was 1 500. The minimum sample weight of child nodes（min_child _weight）was 10. The ratio of feature sampling(colsample_bytree) was 0.8. In the learning target parameters, the objective function selected binary logistic regression（binary: logistic），and the evaluating indicator was the average accuracy（map）. The rest parameters kept

the default values. The results are shown in Table 3 and the values are the average of five experiments.

Table 3   Results of five experiments under different model

	$V_{ACC}(\%)$	$V_{TPR}(\%)$	$V_{SPC}(\%)$	AUC(%)	Time(s)
KNN	88.71	24.78	98.95	61.87	256
BP	80.79	65.10	83.26	74.18	244
RF	91.34	48.87	97.95	73.41	308
XGBoost	84.41	80.68	85.02	82.85	233

It can be seen from Table 3 that the average V_{ACC}, V_{TPR}, V_{SPC} and the time of KNN model are 88.71%, 24.78%, 98.95% and 256 s, respectively. The average V_{ACC}, V_{TPR}, V_{SPC} and the time of the BP neural network model are 80.79%, 65.10%, 83.26% and 244 s, respectively. The mean V_{ACC}, V_{TPR}, V_{SPC} and the time spent on random forest model are 91.34%, 48.87%, 97.95% and 308 s, respectively. Finally, the average V_{ACC}, V_{TPR}, V_{SPC}

and the time spent on XGBoost model are 84.41%, 80.68%, 85.02 and 233 s, respectively. It can be seen that the overall performance of XGBoost classifier is better than that of the other three classifiers. In particular, its sensitivity is far beyond the other three classifiers. When the proportion of positive and negative samples in the data set is not balanced, the reference values of V_{ACC} and V_{SPC} will be reduced. AUC is more capable of representing the performance of the classifier. For example, when the ratio of positive and negative samples is 99：1, in this case, a classifier only needs to determine all the samples to be positive, thus V_{ACC} can reach 99%. Figs. 5－8 are ROC curve graphs of KNN model, BP neural network, random forest model and XGBoost model, respectively. The AUC values of these models are 61.87%, 74.18%, 73.41%, 82.85%, respectively.
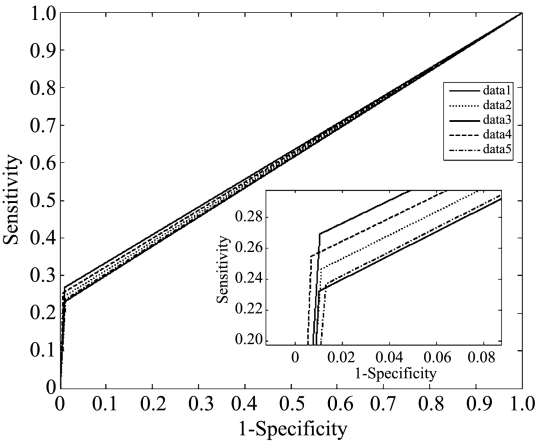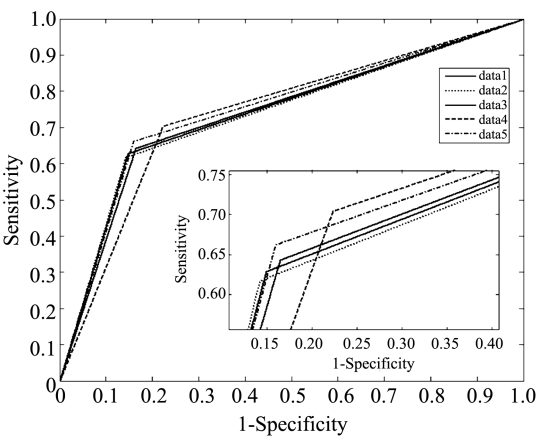

Fig. 5   ROC curve of KNN model
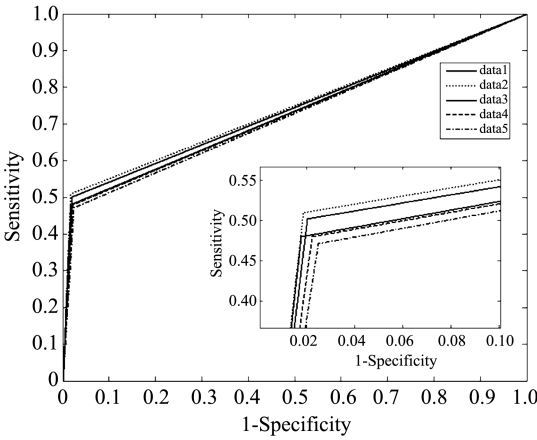

Fig. 6   ROC curve of BP neural network model

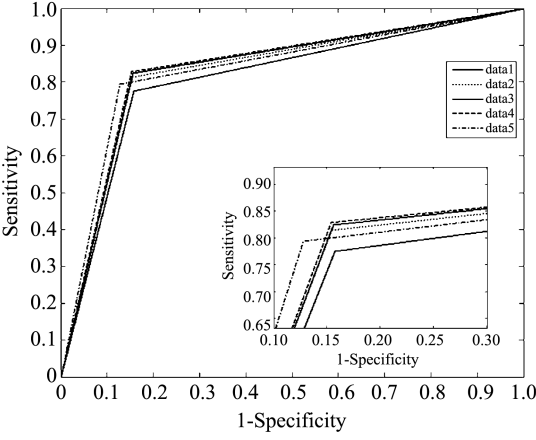**Fig. 7   ROC curve of random forest model**

Fig. 8   ROC curve of XGBoost model

3.4   Analysis

From the above charts, it can be seen that the

XGBoost model outperforms the KNN model, BP neural network and random forest model in both balanced data set and imbalanced data set.

Especially，from balanced data set to imbalanced data set，classification effects of KNN model and random forest model decrease significantly. While the classification effect of XGBoost model does not decrease but slightly improves. Furthermore，the time spent on the XGBoost model is significantly loss than that on the other three models.

Compared with the single algorithm of KNN，XGBoost algorithm is one of the integration algorithm. It combines multiple classifiers and classifies them，showing better generalization ability and classification effect. BP neural network is an excellent classifier and has been widely used. But BP neural network requires strict parameters and has some disadvantages，such as slow convergence speed，ease of falling into local minimum point and difficulty of getting global optimal solution. Random forest and XGBoost are integrated algorithms. Although they all use column sampling to reduce overfitting and computation，there are differences elsewhere between them. Random forest adopts bagging thought，but XGBoost adopts boosting thought. Bagging uses sampling with reset but boosting sampling is based on error rate. Moreover，XGBoost takes the complexity of the model as the regularization term of the objective function to optimize the model. Therefore，the classification performance of XGBoost is better than that of random forest.

In terms of running time，generally speaking，a single classifier consumes less time than an ensemble classifier，but the XGBoost classifier uses less time than KNN classifier and BP neural network. Compared with the random forest，which is also the ensemble classifier，the XGBoost classifier is far beyond the random forest. Compared with the other three models，the advantage of the XGBoost classifier in time mainly attributes to the following two aspects.

1) The XGBoost algorithm performs second-order Taylor expansion of the cost function，and uses the first- and second-order derivatives to converge to the global optimum at the fastest speed.

2) One of the most time-consuming steps of tree learning is to rank the values of the features to determine the best segmentation points. Before training，the XGBoost algorithm sorts the data in advance，and then saves them as a block structure. The structure is used repeatedly in the subsequent iterations，which greatly reduces the amount of computation.

4　Conclusion

In today's world，electricity has been widely used. The development of power big data will accelerate the development of power industry and innovation of business model，achieving the transformation from production-oriented to data-oriented. This study utilizes big data technology，based on different time scale features and correlation coefficient extraction，by using XGBoost classifier to identify abnormal users. Compared with the KNN model，BP neural network model and random forest model，the XGBoost model has achieved good results. In order to deal with the huge amount of data in the power industry，the distribution of XGBoost algorithm will be the next step of our research.

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］ Chen W，Chen Y，Qiu L，et al. Analysis of anti-stealing electric power based on big data technology. Journal of Electronic Measurement & Instrumentation，2016，30(10)：1558-1567.

［3］ Jian F J，Cao M，Wang L，et al. SVM based energy consumption abnormality detection in AMI system. Electrical Measurement & Instrumentation，2014，(6)：64-69.

［4］ Qiu S，Yang H. Sparse Bayesian regression and its application to anomaly detection of harmonic current. In：Proceedings of the CSU-EPSA，2017，(5)：0104-0107.

［5］ Zhuang C，Zhang B，Jun H U，et al. Anomaly detection for power consumption patterns based on unsupervised learning. In：Proceedings of the CSEE，2016，36(2)：379-387.

［6］ Zhou L，Zhao L，Gao W. Application of sparse coding in detection for abnormal electricity consumption behaviors. Power System Technology，2015，39(11)：3182-3188.

［7］ 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.

［8］ Chen T，Guestrin C. XGBoost：a scalable tree boosting system. In：Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，San Francisco，2016：785-794.

［9］ Aler R，Galván I M，Ruizarias J A，et al. Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. Solar Energy，2017，150：558-569.

［10］ Ye Q Y，Rao H，Ji M S，et al. Sales prediction of stores

based on XGBoost algorithm. Journal of Nanchang University，2017，(3)：275-281.

[11] Rui M A，Xie Z，Zhou P，et al. Data mining on correlation feature of load characteristics statistical indexes con-

sidering temperature. In：Proceedings of the CSEE，2015，35(1)：43-51.

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

基于 XGBoost 算法的异常用户识别

宋晓宇，孙向阳，赵　阳

（兰州交通大学 电子与信息工程学院，甘肃 兰州 730070）

摘　要： 电力行业是国民经济中的基础性产业，日益严重的电力异常行为给国家经济造成了巨大损失。XGBoost(eXtreme gradient boosting)算法用于异常客户的识别。首先，对原始数据进行清洗；然后，从不同角度构建用户用电特征；最后，使用 XGBoost 分类器分别在平衡样本集和非平衡样本集下进行异常客户识别。与之对比，在相同特征下，分别使用 KNN(k-nearest neighbor)分类器、BP(back-propagation)神经网络分类器和和随机森林分类器在这两个样本集下进行异常客户识别。实验结果表明，XGBoost 分类器有更高的识别率和更快的运行速度，特别是在不平衡数据集下，性能改进尤为明显。

关键词： 用户识别；用电特征；XGBoost；随机森林