Research on complaint prediction based on feature weighted Naive Bayes

Jian Zheng, Wei Yang\textsuperscript{1}, Changchun Wang, Dandan Jiang, Dan Wang, Qi Yang and Yijiao Zhang

State Grid Fuxin Electric Power Supply Company, Fuxin, China

\textsuperscript{1}E-mail: 2927865106@qq.com

Abstract. Complaint prediction is a hot topic in current research. Due to complaint prediction being random, the accuracy of complaint prediction is not high. To solve this problem, this paper proposes a complaint prediction method based on feature weighted Naive Bayes. According to rough set theory, we first calculate the importance of each conditional attribute, combine the importance of attribute as the weight with the naive Bayes classifier to form a weighted naive Bayes classifier. We use multiple classifiers to make predictions, the latter classifier performs iterative learning on the basis of the previous classifier, and finally all classifiers are given different weights to make decisions. The experimental results show that the proposed method effectively combines the weighting method and the Naive Bayes method to achieve reliable prediction of complaints.

1. Introduction

As the basic industry of the national economy, the stability of the power industry not only affects public security and economic development, but is also closely related to the quality of life of the people. At present, people's living standards are constantly improving, and power customers have higher and higher requirements for electrical energy. In the process of power supply services, a little careless handling may cause complaints from power customers and even develop into public opinion incidents. At present, most electric power companies only understand customers' demands in a passive way. They realize the seriousness of the situation after customers' complaints. At this time, they try their best to make up for their mistakes. Even if they spend a lot of human and material resources to solve customers' demands, they can only achieve half the result with twice the effort, and have caused bad perception to customers. Therefore, how to reverse the current passive situation, from passive understanding of customer demands to active early warning control has become particularly important [1].

The complaint early warning problem is essentially a classification problem, which can be implemented using machine learning algorithms. The commonly used machine learning methods for classification include random forest algorithm [2], decision tree [3], neural network [4], and naive Bayes algorithm [5] and so on. The complaint problem is a small probability event, the complaint data sample is relatively small, but the neural network requires a large number of training sets, which may cause the problem of insufficient prediction accuracy; The random forest algorithm is composed of a large number of decision trees, and the final result is determined by voting on the learning results of each decision tree. The random forest algorithm is fast in training and can process high-dimensional
data, but the amount of data needs to be large. For the case of small amount of data, the Naive Bayes algorithm is relatively suitable.

Naive Bayes can overcome the shortcomings of small data volume and insufficient sample information, but it treats all attributes as independent, in the modeling process some important data information will be lost, so it has certain limitations. At present, many researchers are committed to the improvement of the model, which can be roughly divided into two categories. One is to relax the limitation of the independence between attribute variables and consider the local correlation; the other is to still use the improved model of Naive Bayes, but weighting from different angles to achieve the effect of improving classification [6].

Yingkui Guo and Lihua Lu proposed a naive Bayes method based on feature weighting to identify users of stealing electricity. The weight coefficient of each attribute feature is further calculated by the feature information value, eliminating the influence of the inconsistency of the influence degree of each attribute on the performance of the model [7]. Ming Li, Kefeng Li proposed a novel causality-based attribute weighting method to establish the weighted NBC called IFG-WNBC, where causal information flow (IF) theory and genetic algorithm (GA) are adopted to search for optimal weights [8]. In [9] used a statistical feature weighting technique and proposed a new class-specific deep feature weighting method for Multinomial Naïve Bayes text classifiers. Liangjun Yu, Shengfeng Gan, Yu Chen, Meizhang He proposed the improved model is called correlation-based weight adjusted naive Bayes (CWANB) [10]. Extensive experimental results show that CWANB outperforms NB and some other existing state-of-the-art attribute weighting approaches in terms of the classification accuracy [10]. By introducing the regularization term, in [11], the authors proposed namely regularized naive Bayes (RNB), could well capture the data characteristics when the dataset is large, and exhibit good generalization performance when the dataset is small.

Therefore, this paper proposes a naive Bayes algorithm based on feature weighting, which combines feature weighting and naive Bayes to establish a complaint prediction model based on feature weighted naive Bayes to achieve the purpose of accurately predicting complaints.

2. Feature weighted Naive Bayes algorithm

In order to improve the accuracy of the naive Bayes algorithm, this paper combines the naive Bayes algorithm and the weighted method to form a feature-weighted naive Bayes algorithm, combines the advantages of the two methods to achieve a reliable prediction of complaints.

2.1. Naive Bayes

Naive Bayes is a classification method based on Bayes theorem and the assumption of independence of characteristic conditions [12]. The principle of model classification is as follows:

First point: The data sample is represented by an n-dimensional feature vector \( X = \{ x_1, x_2, \ldots, x_n \} \) which describes the n-dimensional quantity with n attribute samples.

Second point: It is known that there is an m class \( C_1, C_2, \ldots, C_m \). When there is a data sample \( X \) with an unknown label, the classification algorithm predicts that \( X \) belongs to the category with the highest posterior probability under the condition of \( X \). In other words, the condition for the naive Bayes classification algorithm to assign unknown samples to class \( C_i \) is:

\[
P(C_i | X) > P(C_j | X), 1 \leq j \leq m, j \neq i
\]  

(1)

Third point: To maximize \( P(C_i | X) \), the class \( C_i \) with the largest \( P(C_i | X) \) represents the largest posterior hypothesis. According to Bayes’ theorem:

\[
P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)}
\]  

(2)

Fourth point: If the data set has more attributes, the time complexity in calculating \( P(X | C_i) \) may be very large. In order to reduce the time complexity of calculating \( P(X | C_i) \), the independence
between attributes can be assumed first. Assume that each attribute is independent of each other, that is, there is no dependency between the attributes.

The Naive Bayes model has a solid theoretical foundation of statistics and a relatively stable classification efficiency, so this paper chooses the Naive Bayes as the basis to predict complaints.

2.2. Feature weighted Naive Bayes

The Naive Bayes model needs to estimate few parameters, is not sensitive to missing data, and the algorithm implementation is relatively simple. However, the current naive Bayes algorithm also has some shortcomings in classification prediction:

The premise of Naive Bayes theory is to assume that the attributes are independent of each other, but this assumption is often invalid in practical applications. There may be a lot of redundancy between attributes, which will affect the classification efficiency of the Naive Bayes model. When the correlation between attributes is small, the naive Bayes model will show better classification performance [13]. According to the above shortcomings of the current naive Bayes model, corresponding optimizations are made from the independence of attributes and classification decision-making. The specific improvement ideas are as follows:

According to the degree of influence of different attributes on the complaint classification results, different weights are assigned. Therefore, this paper expands the naive Bayes classifier to a weighted naive Bayes classifier. The constructed feature-weighted naive Bayes classification model is:

$$C_{FNBC} = \arg \max \sum_{i=1}^{n} \omega_i \cdot p(c) \prod_{i=1}^{n} p(a_i | c)$$

Although the Bayesian algorithm can make more scientific judgments in the field of classification, subjective probabilities must be used for some data in the calculation process, which ultimately leads to deviations in the classification results.

The AdaBoost algorithm [14] provides a framework to enhance the learning ability of the algorithm through iterative training of misclassified samples. Thus, the weak learner with slightly lower prediction accuracy is upgraded to a strong learner with high prediction accuracy.

The basic idea of the promotion method is to increase the weight of the wrong sample of the previous weak classifier, so that the next weak classifier can learn this "residual", which is equivalent to handing over the problem to multiple weak classifiers to divide and conquer. The classifier learns the data that the previous classifier is not good at, and finally voted together. The classification results of the classifier with a low error rate during voting should account for a larger proportion.

Re-weighting is used to increase the weight of the method, that is, the same weight is initially added to each sample. When calculating the Bayesian estimation of conditional probability, it is no longer a simple addition of the number of sample features. It is the addition with the weight of each sample.

Based on the above improvement ideas, a weighted AdaBoost-NBC algorithm is proposed for power complaints. Figure 1 shows the flow of the weighted Bayes algorithm.

---

**Figure 1.** Weighted-Adaboost-NBC algorithm model and algorithm operation process.
For the weighted-AdaBoost-NBC algorithm proposed in this paper, first calculate the importance of each conditional attribute according to the rough set theory, and combine the attribute importance as the weight with the naive Bayes classifier to form a weighted naive Bayes classifier.

The specific execution steps of the algorithm are:

First point: Calculate the importance of each conditional attribute as the attribute weight according to the rough set method, and construct the weighted naive Bayes classifier proposed above.

Second point: Model initialization, randomly select \( n \) sets of training data from the complaint data set, and initialize the weight \( D(i) \) of the training data

\[
D(i) = (\omega_1, \omega_2, \ldots, \omega_k) = \left( \frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N} \right)
\]

(4)

Third point: Perform iteration, train a weighted naive Bayes classifier. Each sample is attached with an initial weight. When calculating the conditional probability, the number of occurrences of the sample is not added, but the weight value added by the sample is added. After each round of training, the weight of the sample must be updated. Calculate the weight of the weak classifier on the final classifier (\( \alpha_t \) is the weight of the weak classifier).

\[
\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right)
\]

(5)

\( \varepsilon_t \) is the sum of the weights of the misclassified samples.

For samples that are misclassified, the weights are updated as follows: \( \omega_{t+1}(i) = \omega_t(i)e^{\alpha_t} \).

For samples that are correctly classified, the weights are updated as follows: \( \omega_{t+1}(i) = \omega_t(i)e^{-\alpha_t} \).

Fourth point: Finally, after \( T \) rounds of training, the weak classifiers are combined into a strong classifier \( h(x) \) according to the weight \( \alpha_t \). Assume that each attribute is independent of each other, that is, there is no dependency between the attributes.

The classification strategy of the strong classifier is: each weak classifier is attached with its own weight for voting to determine the final category of the sample.

3. Naive Bayesian network construction based on feature weighting

To construct a feature weighted Naive Bayes model, it is first necessary to perform data preprocessing on historical complaint data and basic user's data. Data pre-processing in this article is mainly divided into data conversion, data integration and data cleaning.

3.1. Data conversion

User's historical complaint data and user's basic data mainly include text data and digital data. Text data mainly refers to related data describing power supply areas and complaints. This article uses One-hot discretization coding for processing. One-hot encoding, also known as one-hot encoding. The method is to use N-bit status registers to encode N states, each state has its own independent register bit, and at any time, only one of them is valid. Suppose we have four samples (rows), and each sample has three features (columns). As shown in Table 1, the data features in Table 2 can be obtained by using one-hot encoding.

| Sample | gender | type of electricity | satisfaction of the return visit |
|--------|--------|---------------------|---------------------------------|
| User 1 | 1 (male) | 4 (Commercial electricity) | 3 (Satisfied) |
| User 2 | 2 (female) | 3 (Electricity for urban residents) | 2 (Ordinary) |
| User 3 | 1 (male) | 2 (Electricity for agricultural production) | 2 (Ordinary) |
| User 4 | 2 (female) | 1 (Electricity for rural residents) | 1 (Dissatisfied) |
Table 2. Features after One-hot encoding.

| Sample | gender | type of electricity | satisfaction of the return visit |
|--------|--------|---------------------|----------------------------------|
| User 1 | 01     | 1000                | 100                              |
| User 2 | 10     | 0100                | 010                              |
| User 3 | 01     | 0010                | 010                              |
| User 4 | 10     | 0001                | 001                              |

Digital data mainly refers to data such as the time of complaint and user's power consumption information. For power consumption information, it is converted into digital data by using the standard deviation method, and the correlation between the complaint time data and the power consumption information data is analyzed, and mining potential relevance.

The above-mentioned work is mainly to complete the data complaint work order, and the purpose is to convert the complaint data such as text that cannot be directly recognized by the computer in the work order into data that can be recognized by the computer.

3.2. Data integration and cleaning

Data integration, perform data integration on basic data and complaint data that have been transformed. Data integration is to integrate the user's historical complaint data, such as the time of complaint, the type of complaint, and the attributes of the user's basic attributes such as electricity consumption and region, as a database to participate in subsequent modeling.

This article selects the complaint data of a certain power company, and selects some of complaint users’ data as shown in Table 3.

Table 3. Some of complaint users’ data.

| Account numbers | gender | type of electricity | age | Electricity consumption in January 2020 | Number of complaints per year | 10kv line information(*) | satisfaction of the return visit |
|-----------------|--------|---------------------|-----|-----------------------------------------|--------------------------------|--------------------------|----------------------------------|
| 0250183254      | male   | Commercial electricity | 37  | ***                                    | 126.60                         | 1                        | ***                              | Satisfied                        |
| 0250373868      | male   | Electricity for urban residents | 45  | ***                                    | 170.42                         | 1                        | ***                              | Ordinary                         |
| 0250657779      | female | Electricity for agricultural production | 35  | ***                                    | 55.89                          | 1                        | ***                              | Ordinary                         |
| 0250726604      | female | Electricity for rural residents | 56  | ***                                    | 152.14                         | 1                        | ***                              | Dissatisfied                     |

Note: the data with (*) were not shown due to privacy reasons.

Data cleaning, data cleaning is mainly to delete invalid data in user's historical complaint data and user's basic data and add missing data. The main methods in this article are:

In response to the lack of data complaints from all power users, remove the entire row of data where most of the data is empty. In response to the problem of missing attributes of power user's complaints, remove attributes with high missing rates, and calculate standard deviations for numerical attributes, and remove features with small standard deviations.

Fill in the data and attributes with fewer missing values. This paper uses the method of pre-filling-curve clustering-secondary filling to complete the missing electricity data. First pre-fill the data to
calculate the similarity between the curves. On this basis, for each electricity consumption curve to be completed, find the most similar k curves to form a data matrix, and then apply the matrix filling method to fill in the missing data twice.

In the process of data processing, you will find that the value of one or more fields of some records is very different from the value of other records, or is not meaningful at all, then these records are considered to be abnormal data. When abnormal data are found, the average value of normal data at multiple adjacent times is used to replace the abnormal value. The average value is calculated as shown in the following formula:

\[ x_t = \frac{1}{m} \sum_{i=t-m}^{t-1} x_i \]  

(6)

Among them, \( x_t \) represents the value corresponding to the t-th time, and m is the number of records of normal data at the previous adjacent time. This method comprehensively considers the information of multiple recent moments, weakens the influence of other factors, can fill in the missing values more reasonably, and retain the continuity of the complaint data.

The value range of different numerical fields in the electricity user data may be quite different. For example, the monthly electricity consumption is generally based on hundreds of digits and thousands of digits, while fields such as the number of complaints is mostly single digits or tens of digits.

Therefore, it is necessary to carry out standardization processing, so that the characteristic field data of different value ranges are in the same magnitude range for better modeling and analysis. In view of the large difference in the value range of electricity user data, and the maximum and minimum values of each feature field can be obtained, Min-Max standardization is used to perform data standardization processing on numerical field data. Min-Max standardization is shown in the following formula:

\[ \frac{x_i - \min(x_j)}{\max(x_j) - \min(x_j)} \]  

(7)

3.3. Feature extraction

This paper analyses the pre-processed user historical complaint data and basic user data, and extracts the characteristics of complaint.

Complaint characteristics are important factors that describe the possible causes of complaints. The accuracy of prediction depends largely on how well the characteristics of complaints are extracted. Through the pre-processing of user's historical complaint data and user's basic data, factors such as complaint type, complaint time, region, power consumption, gender, age, station number, industry, power consumption type, and complaint playback satisfaction can be initially extracted as factors. Characteristics of complaint. Through the analysis of user's historical complaint data and user's basic data, it is shown that the user's electricity consumption characteristics, age, gender, and time of complaint are of great importance to the complaint, so they are regarded as the characteristics of the complaint. Among them, the user's electricity consumption characteristics are extracted and transformed through electricity consumption data. The data pre-processing and complaint feature extraction process is shown in Figure 2.

Through data preprocessing and feature extraction, unnecessary features such as station area number, 10kv line information, local work order number, city, county, district, etc. are deleted. Then information such as the community name is transformed into the user’s economic characteristics through the housing price ratio, with a value range of 0.5-2. Feature extraction features are transformed into digital features, for example, male is 1, female is 0, age is transformed into 0-9, and the complaint category is represented by category code 1-5.

After data pre-processing and feature extraction, the user's complaint prediction features include: age, gender, economic level, power consumption characteristics, number of historical complaints.
3.4. Establishment of feature-weighted Bayesian model

The establishment of a feature-weighted Bayesian model is divided into two steps: model training and model prediction.

We establish a feature-weighted Bayesian model on the basis of data preprocessing and feature extraction, and then the K-fold cross-validation method is used to validate and optimize the established feature-weighted Bayesian model.

Model prediction inputs the preprocessed data to be classified, predicts through the established feature-weighted Bayesian model, outputs the classification results, and then periodically updates the training set and feeds back the prediction results. Give the model, let the model continuously iteratively optimize. Figure 3 shows the training and prediction process of the model.

4. The weighted Naive Bayesian algorithm application results

In order to ensure that the coverage of the experimental data is as wide as possible and fully consider the complaints of electricity customers, a number of work order information is randomly selected from the existing user complaint data, and the basic information of the users is correlated to obtain electricity consumption information, which is recorded as set D. Based on the sample set D, the K-fold cross-validation method is used to divide the training set and the test set.
This paper implements the construction of Bayesian network based on Python. Based on sample data, we obtain user's complaint-related information, analyze the basic information and complaint information of complaining users, and obtain the associated attributes of each attribute node to form a preliminary Bayesian network. After gradual optimization, the optimal Bayesian network is obtained. At the same time, the test set is used for verification, and the comparison between the test verification result and the actual situation is given. The experiment process is as follows:

First point: Determine the attributes of each node in the Bayesian network, record it as P, and use the complaint sensitivity attribute as the output attribute;
Second point: Calculate the probability value of each attribute node in the sample data set;
Third point: Calculate the conditional probability based on the attribute value of each node as the prior condition, including the prior probability and the posterior probability value;
Fourth point: Set a threshold for the posterior probability of each node, and if the threshold is exceeded, connect the corresponding nodes to obtain the initial Bayesian network;
Fifth point: Iteratively optimize the initial Bayesian network.

From Figure 4, we can get the final analysis of Bayesian network's complaint sensitivity.

![Figure 4. Bayesian network model of complaint sensitivity.](image)

After obtaining the Bayesian network of complaint sensitivity, the K-fold cross-validation method is used to verify, and the verification is performed according to 10 different test sets, and the average of the 10 test results is used as the final verification result, based on the Bayesian network. The predicted power outage complaints are shown in Figure 5, with an average error of 7.95% and an

![Figure 5. Complaint prediction based on feature weighted Bayesian network.](image)
overall test accuracy rate of 92.05%. The prediction accuracy of the improved feature-weighted Bayesian model is 10.56% higher than that of the basic Bayesian model. The experimental results show that the Bayesian prediction model based on feature weighting proposed in this paper has good accuracy.

Using the feature-weighted Bayesian model proposed in this paper to predict the monthly complaint volume in a certain area of Liaoning Province, and compare it with the prediction result of the naive Bayesian prediction model. The result is shown in Figure 6.

![Figure 6](image)

**Figure 6.** Comparison of predicted results and true values.

It can be seen from Figure 6 that the prediction effect of the model proposed in this paper is closer to the true value than the naive Bayes model, which verifies the accuracy and practicability of the power user's complaint prediction model based on the feature-weighted Bayes algorithm. The model in this paper can realize the forecast of the number of complaints of the power company in the future, and has a certain promotion and application value.

This article analyzes the complaint information of a certain power company. Predicting user complaints through the weighted Bayes algorithm plays an important role in responding to complaints in advance and reducing industry losses. However, the current study generalizes all complainants, and the randomness of complaints is relatively large. There is no further difference between the results of the current study. The complaint level of the complainant, generalized early warning of the predicted users will lead to a waste of manpower and financial resources, and then the subdivision of the complainant will be carried out to predict the complaint probability of the complainant, and the power user will be evaluated based on the prediction result. Classification, according to different levels of users to specify different response measures, while ensuring the company's image, further reducing the cost of responding to complaints, which is an important help in maintaining the image of the power industry.

5. Conclusions

The Bayesian prediction model based on feature weighting proposed in this paper is based on the relevant theory of Bayesian classification, combined with feature weighting method, to study the power user's complaint prediction model. By obtaining complaint data from a certain area in Liaoning Province, the proposed feature Bayesian network model was verified by experiments. The experiment shows that the feature-weighted Bayesian network model studied in this paper has an accuracy of 92.05% in predicting power user complaints, which is 10.56% higher than the basic Bayesian model. It has a good application prospect in power user complaint prediction.
Acknowledgement
This paper was partially supported by Science and Technology Project of State Grid Corporation of China (2021YF-29).

References
[1] Chen Huajun 2015 Big data revolution-exploration of the application of big data in China Southern Power Grid China Electric Power Enterprise Management 17
[2] Zhu Longzhu, Gong Lihua, Liu Kunpeng, Yang Jing and Zhao Qiang 2018 Complaint early warning model optimization method based on random forest algorithm Electric Power Information and Communication Technology 16 8
[3] Arigi Awwal Mohammed, Park Gayoung and Kim Jonghyun 2021 An Approach to Analyze Diagnosis Errors in Advanced Main Control Room Operations Using the Cause-Based Decision Tree Method Energies 14 13
[4] Dudek Grzegorz 2021 Short-Term Load Forecasting Using Neural Networks with Pattern Similarity-Based Error Weights Energies 14 11
[5] Elhadi Aker, Mohammad Lutfi Othman, Veerapandiyan Veerasamy, Ishak bin Aris, Noor Izzri Abdul Wahab and Hashim Hizam 2020 Fault Detection and Classification of Shunt Compensated Transmission Line Using Discrete Wavelet Transform and Naive Bayes Classifier Energies 13
[6] Zhong Xincheng 2019 Research on Early-warning Classification of Naive Bayesian Academic Information Based on Feature Weighting Journal of Shanxi Datong University (Natural Science Edition) 35 02
[7] Yingkui Guo and Lihua Lu 2021 Research on Recognition and Classification of User Stealing Detection Based on Weighted Naive Bayes 2021 International Conference on Control Science and Electric Power Systems (CSEPS) 23
[8] Ming Li and Kefeng Liu 2019 Causality-Based Attribute Weighting via Information Flow and Genetic Algorithm for Naive Bayes Classifier IEEE Access 7
[9] Shufen Ruan, Hongwei Li, Chaojun Li and Kunfang Song 2020 Class-Specific Deep Feature Weighting for Naive Bayes Text Classifiers IEEE Access 8
[10] Liangjun Yu, Shengfeng Gan, Yu Chen and Meizhang He 2020 Correlation-Based Weight Adjusted Naive Bayes IEEE Access 8
[11] Shihe Wang, Jianfeng Ren and Ruibin Bai 2020 A Regularized Attribute Weighting Framework for Naive Bayes IEEE Access 8
[12] Zhao Haixia, Li Yun and Shi Hongbo 2020 Research on Weighted Naive Bayes Algorithm Based on High-dimensional Data Statistics and Decision 36 08
[13] Jia Xian, Liu Peiyu and Gong Wei 2013 Research on Naive Bayes Intrusion Forensics Based on Improved Attribute Weighting Computer Engineering and Applications 49 07
[14] Qiao Wenyu, Li Ye, Liu Haoyu, Li Yang and Yang Ting 2020 Missing electricity data completion method based on curve similarity and low-rank matrix Electric Power Construction 41 1