Research on multi-factor Identification model of acceptance content of Work order complaint

Shang Huaiying 1, Liu Yan 1, Zheng Angang 1, Zhang Qi 1,2 and Ren Min 3

1China Electric Power Research Institute, Beijing 100192, China
2College of Computer Science, Beijing Institute of Technology, Beijing, Beijing, 100081, China
3State Grid Anhui Electric Power Co., Ltd., Hefei 230061, China
ashanghuaiying@epri.sgcc.com.cn e-mail: shanghuaiying@163.com

Abstract. In this paper, combined with the power business demand, the purpose of this paper is to break the blind area of customers' demand for electricity, so as to improve the management level of users' electricity demand, and realize the mining of work order types of hot complaint business. In order to obtain the information behind the complaint work order, the power customer complaint work order is deeply mined. The weight value of keywords is calculated based on TF-IDF algorithm, and the text classifier model is constructed based on SVM classification algorithm. This study aims to provide differentiated service strategies for different types of power customers by controlling the main problems of current power customer complaints, so as to improve customer satisfaction and loyalty.

1. Introduction
The traditional complaint handling and analysis method is manually classified according to the complaint classification when the complaint is accepted, and the content of the complaint is analyzed one by one. This method has some difficulties, such as inaccurate classification, difficult content analysis and low analysis efficiency. By analyzing a large number of historical complaint work order data and based on natural language processing technology and machine learning technology, this study constructed a text classification model to realize the automatic classification of state Grid company's complaint work order[1-2].

2. Design ideas
This study combs the data of customer service complaint work orders of State Grid, analyzes the main complaint tendency of customers and sorts out the complaint attribution framework. Based on the attribution framework, the responsible departments, professional classification, appeal events and error points of historical complaint work orders are marked. The regular expression technology is used to clean the text, remove greetings, idioms and other words that are not helpful for classification. By building a professional word bank for word segmentation, combined with Chinese natural language word segmentation technology, accurate segmentation of complaint content can be achieved. Regular expression technology is used to clean the text and remove greetings, idioms and other words that are not helpful for classification. By building a professional word bank for word segmentation, combined with Chinese natural language word segmentation technology, accurate segmentation of complaint content is realized. The improved TF-IDF algorithm is used to vectorize the segmented words, and the
SVM algorithm suitable for text classification is used to classify the complaint work order, and the optimal text classification is calculated. The automatic classification of customer complaint content in customer service complaint work order is realized. This study solves the problems of the current manual classification, such as the classification granularity is not fine enough, the human resource consumption is too much, and the classification speed is lagging behind[3-4]. At the same time, based on the results of text classification, multi-dimensional data analysis is carried out to analyze the hot spots of complaints, predict the trend of complaints, timely mine customer demands, grasp customer dynamics, and provide guidance and decision-making for improving customer satisfaction[5-6].

3. Model building

3.1. Data source reference

The model is called through the interface, requesting the data required by the interface. The raw data required is as follows:

| The serial number | The name of the system | Table name | The field names |
|-------------------|------------------------|------------|----------------|
| 1                 | Marketing system       | S_95598_WKST_GW_APPNO |            |
| 2                 | Marketing system       | S_95598_WKST_GW_ACCEPTCONTENT |       |
| 3                 | Marketing system       | S_95598_WKST_GW_BUSITYPECODE |        |
| 4                 | Marketing system       | S_95598_WKST_GW_BUSUSUBTYPE |        |

ACCEPTCONTENT refers to the accepted content, and BUSUSUBTYPE and BUSITYPECODE can be further derived into first-level classification, second-level classification, and third-level classification.

The historical data needed for modeling needs to be manually annotated based on the classification knowledge framework. Because the historical classification knowledge framework is added by different salesmen when classifying work orders according to their own cognition, the classification is very chaotic, the granularity of the same level classification is not consistent, and the logic between different levels of classification is not rigorous. Reasonable construction of classification knowledge framework is the key to the application of text classification, which determines whether the
classification results are helpful to understand the current situation of complaints and improve the service quality. Based on the rich business experience of business experts, combined with the business needs, this paper combs out the classification knowledge framework that meets the actual needs. It is summarized as follows:

Responsible departments: marketing, transportation and inspection, two categories in total.

Professional classification: voltage quality, rush repair service, power supply quality, copy and check, electricity change, business hall, etc. 23 categories.

Appeal events: rush repair, frequent power failure, demand fee, frequent power failure, time-of-use electricity price, billing, new installation, change of table/table box, table, etc. 122 categories.

Error points: low voltage, poor service attitude _K, frequent power failure, retention of electricity charges, refusal/prevarication acceptance _K, refusal to issue invoices for electricity charges, etc., 172 categories.

3.2. Data preprocessing

3.2.1. Clean the text with regular expressions
Text cleaning is essential, which can effectively reduce vocabulary noise, retain more effective text features, obtain better text features, and achieve higher accuracy of the classification model. The specific methods are as follows:

(1) Remove punctuation
Because it does not add any additional information to the text data. Therefore, all symbols removed will help reduce the size of training data and improve model training performance.

(2) Stop using words and remove scarce words
Stopword refers to a word whose information does not help the classification of the model or even leads to some misleading words that should be deleted from the text data. This project collects prepositions, greetings and other words in the complaint work order to create a stop vocabulary database, and removes corresponding words from the complaint text according to the stop vocabulary database, so as to achieve the purpose of cleaning the text.

A scarcity word is a word that exists only in a small number of jobs. Due to its rarity, low-frequency vocabulary has limited improvement on the performance of the model. Therefore, rare words can be replaced with other synonyms to improve the word frequency, or deleted directly to improve the iteration efficiency of the model.

(3) Disambiguation conversion
Disambiguation conversion is carried out for some homophones in the text description, for example, disambiguation conversion is carried out for words such as "paternity" and "long indemnity" to be converted into "compensation".

(4) Idiom removal
Similar to "customer call response" and "request power supply company to deal with it in time" in the text, they are common text of different text categories, with high word frequency but no help for classification, so they should be removed to improve the iteration efficiency of the model.

3.2.2. Construct professional lexicon to improve the accuracy of word segmentation

(1) Participles overview
Text segmentation is a special and important part of Chinese text classification. Since word is the smallest language unit that can be used independently, Chinese text, unlike English, does not have any Spaces between words to indicate the boundary of the word. Therefore, text segmentation is the basic part of Chinese natural language processing technology. The effect of text segmentation determines the performance of subsequent text classification models.
Chinese word segmentation requires the use of Chinese word segmentation packages, such as Jieba, HanLP, etc., which adopt statistical methods based on large-scale training corpus. The words in the complaint work list that are not in the corpus are called unlogged words. Unlogged words usually cannot be correctly separated by the word segmentation package. Power grid is a highly specialized industry, which contains a large number of unregistered words that do not exist in the external corpus, such as "transportation inspection, copy and verification, business expansion", etc. These unregistered words are extremely important for text classification because they are professional words. Failure to correctly divide words will not only affect the word segmentation accuracy, but also affect the model effect.

By constructing professional dictionary manually and combining with word segmentation algorithm, the accurate word segmentation of unregistered words can be achieved. This project analyzed more than 6,000 complaint work orders, and collected nearly 600 unregistered words, such as beanstalk, photovoltaic power generation, personnel violation, urging payment, industry expansion and change, etc. Thus, the accuracy of word segmentation and the classification effect of the model are greatly improved.

3.2.3. The text vector is represented by TF-IDF algorithm
Feature weight is used to measure the importance of a feature item in the document representation or the ability to distinguish. The general method of weight calculation is to use the statistical information of the text, mainly the word frequency, to give certain weight to the feature item. TF-idf method measures the weight of a feature from two indexes of word frequency and inverted document frequency. Word frequency refers to the frequency of words appearing in the work order, and the inverted document frequency is used to measure the proportion of words appearing in all work orders. The higher the proportion of a word appearing in all work orders, the lower the inverted document frequency and the lower the weight of the word, such as "customer, company". Based on the characteristics of the algorithm, TF-IDF algorithm is adopted to optimize the weight assignment mode.

The process of text cleaning, text segmentation and word vectorization is as follows:
3.2.4. Based on information theory, information gain is used to screen effective features

After vectorization, the text vector has more than 10,000 features. Too many features may lead to dimensional disaster and affect model efficiency. Therefore, it is necessary to screen out effective features for retention and remove irrelevant features to improve the model effect.

The traditional way of feature selection is to set upper and lower limits based on word frequency, and remove words with too high or too low word frequency. This method is simple and easy to do, and it does remove some of the noise. But this method is only a borrowing algorithm, its theoretical basis is insufficient. According to information theory, although some features appear less frequently, they often contain more information and are of great importance to classification. For such features, word frequency methods should not be used to exclude them directly from vector features.

Based on information theory, this study adopts information gain method to screen characteristics. The information gain method measures the importance of a feature item according to the amount of information it can provide for the whole classification, so as to determine the choice of the feature item. The information gain of a feature term refers to the difference in the amount of information that can be provided for the whole classification with or without the feature, and the amount of information is measured by entropy. The effect of information gain is very good. Effective features can be retained as far as possible while invalid features can be removed to improve the performance of the model. In this way, more than 1,000 features are retained, reducing the number of invalid features by 90%.

3.2.5. Discrete features are constructed and one-hot processed

The state grid work order has the original first-level classification, second-level classification and third-level classification. Although the classification is not simple enough to support the needs of business use, adding it as a discrete feature into the model is helpful to improve the accuracy of the model.

The three features of first-level classification, second-level classification and third-level classification should be processed by BUSITYPCODE and BUSUSUBTYPE in the original data. During the training, the discrete features are processed one-hot coding.

3.3. Algorithm selection and model construction

The extraction of classification rules ultimately depends on the operation of algorithms, and the selection of algorithms is particularly important. Naive Bayes classifier is often used in traditional methods, but this algorithm takes prior probability as an important parameter. However, customer service complaints have strong seasonality, such as frequent power failure, high proportion of complaints in summer and low proportion of complaints in winter, prior probability will change with time, which is not suitable for naive Bayes algorithm. This project has tried bayes, XGBoost and other algorithms, and finally adopted SVM support vector machine algorithm with good classification performance and suitable for high-dimensional sparse space. The basic idea of SVM is to find a decision plane in the vector space, which can best divide the data points in the two categories. The SVM algorithm should find the decision plane with the maximum class boundary in the training set. Because of this feature, SVM has good classification performance and model generalization ability.
The text data converted into word vectors were input into the SVM algorithm, and k-fold cross verification was adopted to continuously iterate the model. Various parameters of the model, such as loss function and penalty coefficient, were adjusted according to the classification accuracy of the model. Finally, the optimal parameters of the model are determined.

3.4. Iterative tuning of models

3.4.1. Iterative optimization during model training

In model training, the classification accuracy is taken as the evaluation index, and the algorithm is appropriately selected and the model parameters are adjusted in combination with the distribution characteristics of sample data and the spatial-temporal complexity of model operation, so as to achieve the purpose of continuously improving the accuracy. Taking the automatic classification of the department responsible for work order complaints as an example, the iterative optimization process of the project model in the training stage is as follows:

At the beginning of modeling, according to the characteristics of the short text, the word vector method adopts the de-duplication word bag method, and the model USES Bernoulli Bayesian algorithm, with general effect. The analysis reason should be that Bernoulli Bayesian algorithm "does not exist" also as a feature of the feature, not applicable to the current classification task, consider to replace the algorithm.

The replacement algorithm is polynomial Bayesian algorithm, which eliminates the feature that "the word does not exist" is also a feature. After the parameter adjustment, the effect is significantly increased.

Further cleansing the text, through regular expressions, removes more than 4,000 meaningful words, leaving more than 7,000 words, streamlining the model. At the same time, other features are introduced to incorporate the 1-3 classification in the work order into the model. The effect is improved.

The word vectorization was replaced by the TF-IDF method with more information, and multiple TF-IDF variants were tried to select the most appropriate scheme, which improved the model effect.

Using logistic regression, XGBOOST, SVM and other algorithms, through iterative parameters, continue to try the model effect. XGBOOST takes more time, logistic regression performance is poor, SVM works well and takes less time, and SVM algorithm is selected to have the best classification performance.

In the next step, the model was further optimized by adjusting the classification framework, adding a few class samples, combining the business optimization characteristics, exhausting the parameter space and other methods, and the final accuracy rate was 97%.

4. Conclusion

Based on natural language processing technology and machine learning technology, a text classification model was built to improve classification accuracy by analyzing a large number of historical complaint work order data. This paper discusses the design idea, data source, data preprocessing, algorithm selection, model construction and iterative optimization of the model, so as to realize the automatic classification of state Grid company's complaint work orders.

Acknowledgments

This work is supported by Science and Technology Project of SGCC. (Research on new measurement technology supporting energy Internet)

References

[1] Li Jianbin, Zhu Yakui, Fu Liheng. The customer demands analysis and application based on big data [J]. Power Systems and Big Data, 2017, 20(10): 14-17.
[2] Li Dan. Research on Chinese text classification based on Naive Bayes method [D]. Bao Ding. Hebei University, 2011.

[3] Huang Chenghui, Yin Jian, Hou Fang. A text similarity measurement combining word semantic information with TF-IDF method [J]. Chinese Journal of Computers, 2011, 34(05): 856-864.

[4] Zhou Yuan, Liu Huailan, Du Pengpeng, et al. Research on text classification model based on improved TF-IDF feature extraction [J]. Information Science, 2017, 35(05):111-118.

[5] Zhao Xiaoming, Zhang Xueqiang, Cao Lan. Thematic analysis of "big data" and "cloud computing" in power system based on key words [J]. Zhejiang Electric Power, 2016, 35(02): 27-30.

[6] Ruan Guangce. Analysis on web media report differences based on text mining [J]. Information Science, 2012, 30(01): 105-109.