Research on data mining technology in power marketing system

qi meng[1], Xixiang Zhang[2], and jun yang[1]

[1]Guangxi Power Grid Co., Ltd
[2]China Southern Power Grid Guangxi Power Grid Co Ltd

May 31, 2023

Abstract

This article starts with the importance of electric power marketing systems, introduces the technical characteristics of data mining and its application status in electric power marketing systems, thereby providing decision-making basis for the economic operation of power grids. And propose using C5.0 decision tree algorithm to deeply analyze the marketing data of the electric power marketing management system. The original C5.0 decision tree algorithm is improved by introducing information entropy, which improves its classification speed and accuracy. Experimental results on UCI machine learning dataset and power marketing dataset show that the proposed improved C5.0 decision tree algorithm has good classification performance and can meet the classification and prediction requirements in power marketing work.

Research on data mining technology in power marketing system

Qi Meng[1] , Xixiang Zhang*[1] , Jun Yang[1]
China Southern Power Grid Guangxi Power Grid Co Ltd;
*Correspondence: Zhangxixiangi@outlook.com;

Abstract—This article starts with the importance of electric power marketing systems, introduces the technical characteristics of data mining and its application status in electric power marketing systems, thereby providing decision-making basis for the economic operation of power grids. And propose using C5.0 decision tree algorithm to deeply analyze the marketing data of the electric power marketing management system. The original C5.0 decision tree algorithm is improved by introducing information entropy, which improves its classification speed and accuracy. Experimental results on UCI machine learning dataset and power marketing dataset show that the proposed improved C5.0 decision tree algorithm has good classification performance and can meet the classification and prediction requirements in power marketing work.

Keywords: power marketing system; data mining; C5.0 decision tree; information entropy

1. Introduction

The main problem faced by the current application of information technology in the power marketing system is how to extract valuable information from the massive marketing data, and then provide decision-makers with as accurate and detailed quantitative indicators and decision-making data as possible, to improve the management level and guide the economic operation of the power grid.

Using reasonable data mining technology can solve the data analysis problem of power enterprises, provide valuable decision support information for managers, and improve the reliability of power grid operation[1]. Document [2] uses C4.5 decision tree algorithm and proposes a marketing effect evaluation method based on big data mining on power side. Document [3] uses BP neural network to quickly mine the marketing data of electric power enterprises. However, the decision tree algorithm used in the above methods is still ID3 algorithm and C4.5 method developed on its basis. As the latest version of C5.0 decision tree algorithm, its comprehensive performance has been significantly improved.

This paper starts with the importance of power marketing system, introduces the technical characteristics of data mining and its application status in power marketing system, so as to provide decision-making basis for the economic operation of power grid. It also proposes to use C5.0 decision tree algorithm to deeply analyze the marketing data of the power marketing management system. The original C5.0 decision tree algorithm is improved by introducing information entropy to improve its classification speed and accuracy. The experimental results on UCI machine learning dataset and power marketing dataset show that the improved C5.0 decision tree algorithm has good classification performance and can meet the classification and prediction requirements in power marketing.

2. Power marketing system

The power marketing system provides service modules and analysis modules on top of each business module, with key businesses such as industry expansion and installation, power metering, power consumption management, business billing and line loss management as its core. The service module includes telephone service, Internet service and customer center service, focusing on providing various fast and high-quality services for power users; The analysis module includes comprehensive business query, statistics based on historical data, benefit analysis and decision support, focusing on providing timely and accurate decision basis for enterprise leaders[4]. Therefore, all the original data that provide decision-making for the normal operation of the power system can be summed up in the category of marketing data, such as production system planning and design, load forecasting and user feature extraction, economic dispatching, power system fault diagnosis, dynamic security assessment, mining and corresponding processing of abnormal data, etc.

2.1 Data source of marketing system

The massive data of the marketing system is composed of the operation data of the management information system, geographic information system, SCADA system and the real-time information system (load management system, electric energy billing system, distribution transformer detection system, metering verification) of the power grid operation. With the rapid development of information technology in power enterprises,various systems have generated and accumulated a large amount of historical data.

2.2 Data characteristics of marketing system

(1) Lots of data. In the power marketing system, the data is mainly divided into on-site data collected by various devices in real time and a large amount of data generated by various systems of the dispatching center in the process of operation, with many sources of data. In addition, the power system is a large-scale singular nonlinear dynamic large-scale system, which often involves thousands of state variables when describing its characteristics. The traditional processing method is to reduce the dimension or simplify the system, which to some extent affects the accuracy of the final result [5].

(2) Data types are mixed. The marketing system is a standard hybrid system. The dispatching decisions given by the upper layer, such as the dispatching center, are mainly logical operation instructions, while the lower layer control, such as the excitation and speed regulation control of the generator, is mainly continuous.

(3) Poor data quality. In the marketing system, the collected data often contains noise, missing, incorrect and other situations.

(4) High requirements for data. When the system is in an emergency or even collapsed state, it is necessary to make a real-time online fast decision to return the system to normal state.

3. Data mining technology

Data mining, also known as knowledge discovery in databases, is the product of the combination of database development and artificial intelligence technology, and is a new discipline. It integrates technologies such as statistics, pattern recognition, artificial intelligence, machine learning, database and high-performance parallel computing, and promotes people's application of data from low-level simple queries to mining knowledge from data, providing decision support for managers. It has a very broad application prospect, and is one of the most advanced research directions in the field of database and information decision-making in the world.

3.1 Basic concepts of data mining

Data mining is the process of extracting potentially useful information and knowledge hidden in a large number of incomplete, noisy, fuzzy and random data that people do not know in advance[6]. Data mining technology can be used to classify, extract, optimize and integrate the massive data generated by the power marketing system.After storing the data appropriately, relevant indicators can be calculated in real-time to extract the relevant marketing information. The results will play a data support role in the decision-making of the marketing system, and better guide the management and decision-making levels of enterprises to make rapid and scientific marketing decisions in the changing environment.

3.2 Main technologies of data mining[7]

(1) Association rules. Association rule is one of the most active research methods in data mining. It was first proposed by AGREWAL and others for shopping basket analysis. The classic Apriori algorithm was used to mine valuable knowledge describing the relationship between data items from a large amount of data. With the increasing scale of data collected and stored in the database, people are more and more interested in mining related knowledge from these data. Association rules are widely used in the field of data mining, and are suitable for discovering meaningful relationships between data in large data, and then effectively evaluating these association rules to screen out meaningful association rules that users are really interested in.

(2) Classification. Classification is a very important task in data mining. Its purpose is to learn a classification function or classification model (classifier), which can map data items in the database to a certain category in a given category. Classification can be used for prediction. The purpose of prediction is to use historical data records to automatically derive the generalization description of given data, so as to predict future data.

(3) Time series mining and sequence mining. Time series mining is an important research branch in data mining, and it has wide application value. It refers to extracting information and knowledge related to time attributes from a large number of time series data that people do not know in advance, but potentially useful, and used for short-term, medium-term or long-term prediction to guide people's social, economic, military and life behaviors. Sequence mining, also known as sequence pattern mining, refers to the discovery of high frequency subsequences in relative time or other sequences. As a general method and technology, sequence mining was first proposed by AGRAWAL et al., and has become a new research branch of data mining and has been widely discussed.

(4) Clustering. Clustering is to divide data objects into multiple classes or clusters. The principle of division is that the objects in the same cluster have high similarity, while the objects in different clusters differ greatly. Unlike classification, the class to be divided in clustering operation is unknown in advance, and the formation of the class is completely data-driven, belonging to an unsupervised learning method.

(5) Web mining. Web mining can be simply defined as the application of data mining methods to help people extract knowledge from WWW and provide decision support for visitors, site operators and Internet-based business activities, including e-commerce, for various Web data, including Web page content, structure between pages, user access information, e-commerce information, etc.

(6) Spatial mining. Spatial mining is a data mining technology that has been developed in recent years and has broad application prospects. In essence, it is the application of data mining in spatial database or spatial data. In short, spatial mining is to extract implicit knowledge, spatial relationships or other patterns that are stored in the spatial database in a non-visible way from the spatial database for understanding spatial data and discovering the relationships between data (spatial or non-spatial).

4. Application of data mining technology in power marketing system

4.1 Application of association rules in power marketing system

Association rule is one of the main technologies in the research of power marketing data mining at present. It helps decision-makers analyze the characteristics and laws of historical data and current data to predict the future. Reference [8] introduces association rules into the marketing analysis of the power market, uses FP-Growth algorithm to analyze the association rules of the discretized power marketing data, describes the association characteristics between various external factors that affect the power sales and the power sales level, and provides auxiliary decision-making information for the power marketing. Reference [9] proposes that association rules can guide power marketing strategy, project and portfolio management, and then carry out demand forecast, sales and revenue forecast, claim analysis, etc. In the direction of combining association rules with other methods, many scholars have done in-depth research one after another. Reference [10] proposes a fuzzy evaluation method of power marketing target market based on cloud model and correlation analysis method, which provides a simple and feasible method for determining power marketing target market.

4.2 Application of classification in power marketing system

In the medium and long term prediction, in addition to the traditional sequence prediction methods, fuzzy theory [11], expert system[12] and other methods have been applied. The mode classification method proposed in Reference [13] can improve the accuracy of power load forecasting. Reference [14] proposes a short-term load forecasting algorithm that combines decision tree technology with extrapolation method, and applies it to the preparation of daily dispatching plan in Fujian Province, with high forecasting accuracy. Reference [15] reduces the scale of SCADA database and effectively improves the calculation speed when estimating the bad data state of SCADA system by using the method of sub-database formed by classification tree.

4.3 Application of time series mining and sequence mining in power marketing system

Time series mining is considered to be the most classical, systematic and widely used short-term load forecasting method. In short-term load forecasting, neural network is the most widely used and studied. In practical applications, time series mining and neural network are often combined to analyze power marketing data. Reference [16] has made a good summary of this. Wavelet neural network has faster convergence speed than BP neural network, and the application of its improved membership clustering method can improve the prediction accuracy of load large wave days[17].

4.4 Application of clustering in power marketing system

Clustering is mainly applied to power user classification and credit evaluation, bad data correction, load forecasting and classification, transformer fault diagnosis, etc. According to the different attributes of power supply enterprises' customers in various aspects, the reference [18] uses the cluster analysis method to cluster customers into different groups. The decision analysts can analyze the differences between different groups according to the clustering results, and adopt different marketing strategies through the study of the characteristics of the groups, so as to improve the economic benefits of enterprises. In reference [19], according to the characteristics of power customer credit classification, a power customer credit evaluation algorithm based on fuzzy clustering analysis is proposed, and the clustering centers of different customer groups and the customer membership matrix are obtained, which provides a quantitative basis for the characteristic analysis of customer groups.

4.5 Application of space mining in power marketing system

For the power marketing system, in most cases, decision makers need to quickly analyze, diagnose and make correct response in time. Especially under the power market conditions, the correctness of important decisions is inestimable for the development of power enterprises. By integrating the information of power grid operation data, load location distribution data, real-time change data and other multi-objective levels, and using special space technology to comprehensively process them, advanced functions such as equipment tracking, fault location, simulated outage, loss evaluation and optimal scheduling can be realized[20].

5. Power marketing data mining based on C5.0 decision tree algorithm

5.1 C5.0 Decision Tree Algorithm Principle

As a follow-up algorithm developed from ID3 decision tree algorithm, Ross Quinlan proposed that C4.5 algorithm can treat attributes as continuous[21], and proposed two new attributes: separating information and information gain rate, which can be used to generate multi-branch decision trees. The core of C5.0 algorithm and C4.5 algorithm is the same, but many improvements have been made in memory management and other aspects, which are more suitable for commercial applications.

5.1.1 Determination of split attribute

This paper assumes that S represents a training sample set with a number of s samples, including m different categories x_i (i=1, 2,..., m). D represents an attribute of training sample set S and the value range is [1, k]. V_i represents the total number of samples belonging to different types of x_i.

According to the difference of attribute D, the training sample set S can be divided into k small subsets. Si is the ith subset of the above subset, i=1,2,..., k, [S_i] represents the number of samples in the subset. First of all, the calculation formula[22] of information gain Gain (S, D) is shown in formula (1).

$$Gain(S,D) = I\left(s_1, s_2, \cdots, s_{\bar{k}}\right) - E(S,D) \tag{1}$$

Where, E (S, D) represents the weighted sum of entropy of k subsets divided according to attribute D, and I (s_1, s_2,..., s_k) represents the entropy of training sample set S. The calculation formula of I (s_1, s_2,..., s_k) is shown in formula (2).

$$I\left(s_1, s_2, \cdots, s_k\right) = -\sum_{i=1}^{m} p\left(x_i\right) \log_2 p\left(x_i\right) \tag{2}$$

Where, p (xi) represents the probability of occurrence of category x_i, and satisfies the constraint conditions as shown in formula (3)

$$\sum_{i=1}^{m} p\left(x_i\right) = 1 \tag{3}$$

Secondly, the calculation formula of split information item Split_Info (S, D) [23-25] is as follows

$$\text{Split\_}Info(S,D) = -\sum_{i=1}^{k}\left(\frac{[S_i]}{s}\log_2\frac{[S_i]}{s}\right) \tag{4}$$

It can be seen from equation (4) that the split information item Split_Info (S, D) is actually the entropy of training sample set S about attribute D. The smaller the value, the more uneven the result of sample distribution on attribute D. Therefore, the calculation formula of the gain ratio of information GainRatio (S, D) is

$$\text{GainRatio}(S,D) = \frac{Gain(S,D)}{SplitInfo(S,D)} \tag{5}$$

5.1.2 Determination of split attribute

In the process of splitting attribute selection, C5.0 decision tree algorithm will select the attribute with the highest gain ratio and define it as splitting attribute[26]. After determining the split attribute, C5.0 decision tree algorithm will perform the steps of determining the best split point, forming the k branches of the decision tree. When the optimal splitting attribute is a continuous variable, the partition value will be set using the splitting strategy, so that the samples larger than the partition value can be subdivided[27]. A complete decision tree is generated after the determination of the split attribute and the determination of the best split point.

5.2 Improved C5.0 decision tree algorithm

Through the principle analysis of C5.0 decision tree algorithm in the previous section, it can be seen that the decision tree algorithm needs to calculate the information gain ratio GainRatio (S, D) on all nodes according to each attribute in order to support the determination of subsequent split attributes. However, the calculation of information gain ratio GainRatio (S, D) in equation (4) requires logarithmic operation, so the calculation time is long and the accuracy is not ideal[28]. Therefore, this paper improves the original attribute selection method by introducing information entropy. Assuming that the number of positive case attributes is p, and the number of negative case attributes is n, the calculation method of information quantity is as follows (6),

$$\text{Split\_}Info(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \approx \frac{np}{p+n} \tag{6}$$

Information entropy can be deduced as formula (7)

$$\text{Split\_Info}'(S,D) = \sum_{i=1}^{n}\frac{p_i+n_i}{p+n}SplitInfo\left(n_i, p_i\right) \tag{7}$$

And,

$$\text{Split\_}Info\left(n_i, p_i\right) = -\frac{n_i}{p_i+n_i}\log_2\frac{n_i}{n_i+p_i} - \frac{p_i}{n_i+p_i}\log_2\frac{p_i}{p_i+n_i} \tag{8}$$

Substitute Equation (8) into Equation (7) to get Equation (9)

$$\text{Split\_Info}(S,D) = \frac{1}{(n+p)\ln 2}\sum_{i=1}^{n}\left(-p_i\ln\frac{p_i}{n_i+p_i} - n_i\ln\frac{n_i}{n_i+p_i}\right) \tag{9}$$

Because (n+p) ln2 is constant, you can use formula (10) to select node attributes during the repeated cycle,

$$\text{Split\_Info}'(S,D)=\sum_{i=1}^{n}\left(-p_i\ln\frac{p_i}{n_i+p_i}-n_i\ln\frac{n_i}{n_i+p_i}\right) \tag{10}$$

Because both pi/(ni+pi) and ni/(ni+pi) are less than 1

$$\ln\frac{p_i}{n_i+p_i}=\ln\left(1-\frac{n_i}{n_i+p_i}\right)\approx-\frac{n_i}{n_i+p_i} \tag{11}$$

$$\ln\frac{n_i}{n_i+p_i}=\ln\left(1-\frac{p_i}{n_i+p_i}\right)\approx-\frac{p_i}{n_i+p_i} \tag{12}$$

Therefore, the final Split_Info '(S, D) can be calculated using Equation (13),

$$\text{Split\_Info}'(S,D)\approx\frac{n_i p_i}{n_i+p_i} \tag{13}$$

It can be seen from equation (13) that the calculation process only includes simple addition, subtraction, multiplication and division operations, which greatly reduces the calculation time.

5.3 Application of improved C5.0 decision tree in power marketing

5.3.1 Data model design

When applying the improved C5.0 decision tree to the power plant management information system, it is necessary to build the mapping relationship between the database tables of each department with the sales volume as the center. The data model built in this paper is shown in Figure 1.

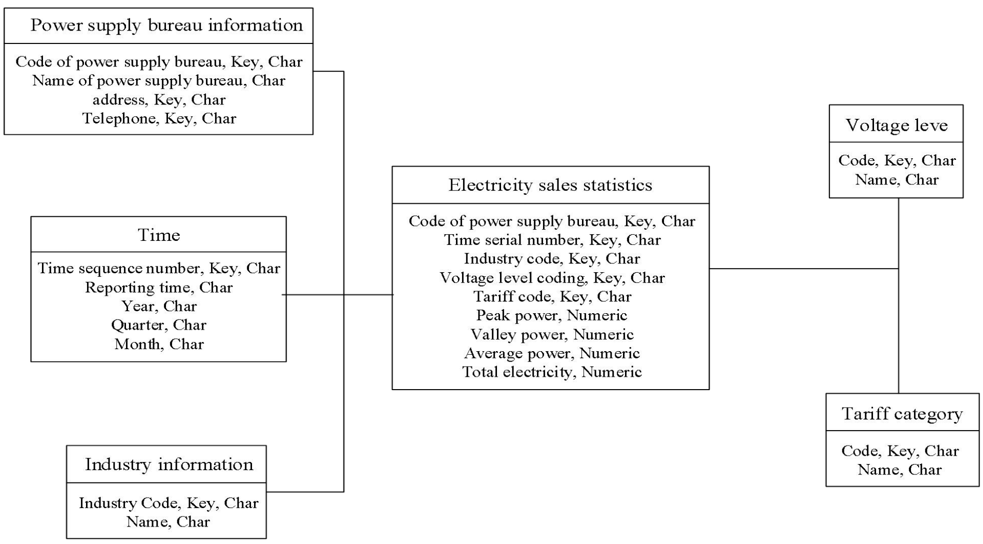

Figure 1 Relationship model of electricity sales

5.3.2 Data mining process

The data mining process of improving C5.0 decision tree in power marketing is shown in Figure 2.

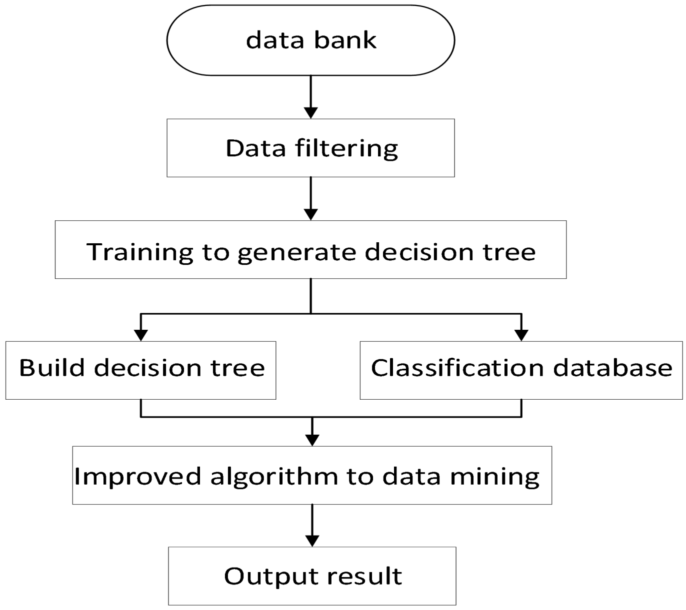

Figure 2 The mining process of Improved C5.0 decision tree algorithm

5.4 Experimental results and analysis

5.4.1 Experimental environment

In order to analyze and verify the video classification method proposed in this paper, concrete experiments are carried out. Experimental hardware environment: the processor is IntelCorei72.2GHz, the graphics and image processing equipment is GTX970 @ 2G video memory, and the memory is 8G. Experimental software environment: Windows7 operating system, MATLAB 7.0 simulation software.

5.4.2 Performance verification of improved algorithm

In order to verify the performance of the proposed improved C5.0 decision tree algorithm, a classification test is carried out with UCI machine learning data set[29]. A total of 2400 samples from 20 groups of small data sets were tested. The training samples are 1000 randomly selected samples, and the rest are test samples. Each group of experiments was repeated 10 times and the average value was taken as the final result. The classification accuracy and time comparison of the standard C5.0 decision tree algorithm and the improved C5.0 decision tree algorithm are shown in Figure 3 and Figure 4 respectively. As can be seen from Figure 3, with the increasing number of test samples, the classification accuracy of the two algorithms is almost the same. As can be seen from Figure 4, with the increasing number of test samples, the classification time of both algorithms has increased, but the improved C5.0 decision tree algorithm requires significantly less time and has higher classification efficiency.

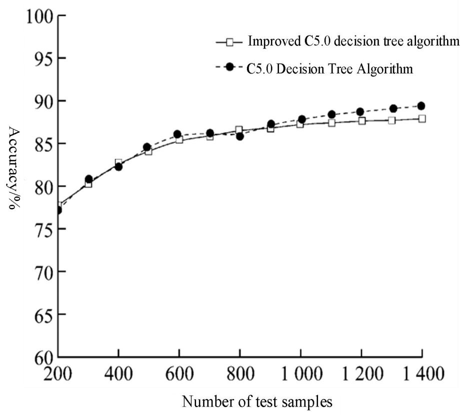

Figure 3 Comparison of classification accuracy of two algorithms

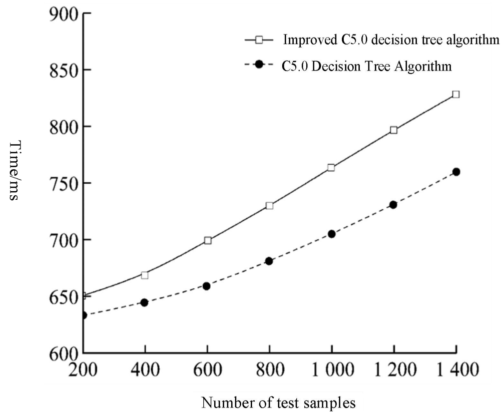

Figure 4 Comparison of classification time of two algorithms

5.4.3 Application results of power marketing

Taking the 2019 marketing data of an electric power enterprise as the test data set, the improved C5.0 decision tree algorithm is analyzed as a practical case. The electricity marketing test data set includes 100 electricity sample data of users in different regions, ages and positions. Select 50 samples randomly as training samples, as shown in Table 1[30].

Table 1 Training Sample Set

age	member	Arrear record	area	Power consumption time
21	no	no	Northern urban area	More than 20 years
19	no	yes	Northern urban area	More than 20 years
40	yes	no	Northern urban area	More than 20 years
58	yes	no	Gaokai District	More than 20 years
37	yes	yes	Southern urban area	Less than 20 years
21	no	no	Gaokai District	More than 20 years
15	yes	no	Southern urban area	Less than 20 years
45	yes	no	Gaokai District	Less than 20 years
46	yes	yes	Gaokai District	Less than 20 years
43	yes	yes	Gaokai District	More than 20 years
45	yes	no	Northern urban area	Less than 20 years
60	no	yes	Gaokai District	More than 20 years

Using the Improved C5.0 decision tree algorithm, select the attribute value with the maximum information gain as the leaf node, cycle the above decision tree execution steps, finally generate the customer classification decision tree and its classification rules, and then verify the classification decision tree model in the test samples of the power marketing data set. In addition, for comparative analysis, BP neural network and ID3.0 decision tree are used to establish classification models respectively. The results of various classification models are compared as shown in Table 2.

Table 2 Comparison of Results of Various Classification Models

Model	Classification accuracy/%	Classification time/s
ID3.0 decision tree	78.9	2.32
BP neural network	84.1	2.71
C5.0 Decision Tree Algorithm	87.6	2.14
Improved C5.0 decision tree algorithm	86.5	1.86

It can be seen from Table 2 that compared with BP neural network and ID3.0 decision tree, C5.0 decision tree and improved C5.0 decision tree algorithm have significantly improved the classification accuracy, C5.0 decision tree is the best (87.6%), and improved C5.0 decision tree algorithm takes the second place (86.5%). In terms of classification efficiency, the improved C5.0 decision tree algorithm is outstanding, only 1.86s. In general, the improved C5.0 decision tree algorithm has the best comprehensive classification performance in the application of timely and accurate customer classification, which can effectively meet the needs of actual power marketing work.

6. Conclusion

The overall application of data mining in the power marketing system is still at a relatively low level. A single mining algorithm is difficult to meet the needs of actual decision-making. Therefore, the mining algorithm should be continuously improved under the influence of various factors. Although there are shortcomings, data mining has a high predictability of potential problems and laws, and has the ability of efficient calculation, monitoring and management. Therefore, it is suitable for solving large-scale nonlinear problems in power marketing systems, and will show its strong vitality and become an important tool for the development of power marketing. This paper proposes a data mining technology for power marketing based on improved C5.0 decision tree algorithm. The original attribute selection method is improved by introducing information entropy, which improves the speed of information gain ratio calculation. In addition, the data in the power plant management information system is mined according to the designed relationship model of electricity sales. The effectiveness and feasibility of the proposed improved algorithm are verified by the data set and the actual case application results. However, the number (dimension) of user attributes in the power marketing training and test sample set is small, and more attributes will be considered later to further verify the performance of C5.0 decision tree algorithm.

6. Patents

Author Contributions: Conceptualization and methodology, Qi Meng; software, validation and writing—original draft preparation, Xixiang Zhang; data curation, review and editing, Jun Yang.

Funding: No Funding

Data Availability Statement: Not applicable.

Conflicts of Interest: no conflict of interest.

References

1. Li Baoshu, Qiu Hongjian, Zhou Xin Design of power load automatic monitoring system based on big data [J] Electronic Design Engineering, 2019, 27 (08): 95-98+103.
2. Liu Jianfei, Li Xiang, Yang Ke, Liu Xiaomeng. Evaluation and analysis of new marketing effects based on multi-dimensional clustering mining of big data on power consumption [J]. Modern Information Technology, 2019, 3 (22): 142-143+14.
3. VICTOR D, KUAMR C. Modular Implementation of Neural network Structures in Marketing Domain Using Data Mining Technique[J]. International Journal of Advanced Trends in Computer Science ＆ Engineering, 2016, 5(1):265-323.
4. Xu Chunling, Jiang Yu, Zheng Xiao'an. Power marketing system based on three-layer architecture of application server [J]. Power Grid Technology, 2002 (07): 71-73.
5. Lu Qiang. Digital Power System [J]. Power System Automation, 2004 (9): 1-4.
6. HAN Jiawei, KAMBER M. Data mining: concepts and technologies [M]. Translated by Fan Ming and Meng Xiaofeng. Beijing: China Machine Press, 2001:185-195.
7. Mao Guojun, Duan Lijuan, Wang Shi, et al. Principles and algorithms of data mining [M]. Beijing: Tsinghua University Press, 2005.
8. Hou Xuebo, Tian Bin, Ge Shaoyun, et al. Application of association rule technology in power marketing analysis [J]. Journal of Power System and Automation, 2005,17 (2): 67-72.
9. Zhang Bin, Sun Zhe. Application of association rule technology in power system [J]. Grid Technology, 2007,31 (1): 64-66.
10. Zhang Ke, Wang Yufan, Yuan Jinsha, et al. Fuzzy evaluation of power marketing target market based on cloud model and correlation analysis [J]. Journal of North China Electric Power University, 2009,36 (4): 30-34.
11. Sun Hongbo, Xu Guoyu, Qin Yihong. Application of fuzzy theory in power load forecasting [J]. Journal of Chongqing University: Natural Science Edition, 1994,17 (1): 18-22.
12. KANDIL M S,EL-DEBEIKY S M,HASANIEN NE. Long-term Load Forecasting for Fast Developing Utility Using a Knowledge Based Expert System[J].IEEE Transactions on Power Systems,2002,17(2):491-496.
13. Feng Li, Qiu Jiaju. Short-term power load forecasting based on power load mode classification [J]. Power Grid Technology, 2005,29 (4): 23-27.
14. Yao Yuzhen, Lin Lijuan. Application of extrapolation algorithm based on decision tree in short-term load forecasting of Fujian power grid [J]. Fujian Electric Power and Electrician, 2007,27 (2): 24-27.
15. Guan Shengli. Applying data mining technology to estimate the bad data status of SCADA system [J]. Electric Power Informatization, 2007,5 (4): 70-72.

16. HIPPERT H S, PEDREIRA C E, SOUZA R C. Neural Networks for Short-term Load Forecasting:A Review and Evaluation[J]. IEEE Transactions on Power Systems,2001,16(1):44-55.
17. Zhang Buhan, Zhao Jianjian, Liu Xiaohua, et al. A short-term load forecasting method based on wavelet neural network [J]. Power Grid Technology, 2004,28 (7): 15-18.
18. Feng Lu, Wang Chengwen, Shen Xiaoliu, et al. Research and design of customer relationship management system for power supply enterprises based on data mining [J]. Power Information, 2007,5 (7): 86-89.
19. Li Bin. Power customer credit evaluation algorithm based on fuzzy clustering analysis [J]. Measurement Technology, 2010,30 (1): 18-21,34.
20. Guo Yingqi, Chen Linling. Application of spatial data technology in 10 kV distribution network dispatching GIS[J]. Computer Application, 2002,22 (8): 91-93.
21. Xu Xuran, Tu Juanjuan. Air quality prediction system based on decision tree algorithm ［J］. Electronic Design Engineering, 2019, 27 (9): 39-42.
22. Cheng Hui, Zhang Rui, Zhang Shike, et al. blackout sensitivity analysis based on improved decision tree［J］. Microcomputer Application, 2020, 36 (3): 144-148.
23. Gong Linming, Li Shundong, Dou Jiawei, et al. RSA encryption scheme against CPA and CCA2 under standard model [J]. Journal of Electronics, 2018, 46 (8): 1938-1946.
24. Xia Bo, Yang Chao, Zheng Kaiwen. Design and implementation of power load forecasting system under big data platform [J]. Automation Instrument, 2018, 39 (6): 81-84.
25. Li Xiu, Lu Nan, Chen Zhou. Research on Android malicious application detection method based on decision tree [J]. Information Technology, 2018 (5): 125-12.
26. SAVE P, TIWAREKAR P, KETAN N, et al. A Novel Idea for Credit Card Fraud Detection using Decision Tree[J]. International Journal of Computer Applications, 2017,161(13):6-9.
27. Ding Bing, Wu Yunping, Li Binya. A parking space recognition method based on C4.5 algorithm [J]. Electronic Measurement Technology, 2015, 38 (8): 64-68.
28. TANHA J, SOMEREN M, AFSARMANESH H.semi-supervised self-training for decision tree classifiers[J]. Inernational Journal of Machine Learning and Cybernetics,2017,8(1):355-370.
29. TAYEFI M, ESMAEILI H,SABERI Karimian M,et al. The application of a decision tree to establish the parameters associated with hypertension[J].Computer Methods and Programs in Biomedicine, 2017,139:83-91.
30. Wu Renbiao, Huang Cheng, Wang Xiaoliang, He Weikun, Liu Shanliang. Analysis of the impact of MDI training sample set construction on radar detection of rotor drone classification [J]. Signal Processing, 2021,37 (06): 1017-1033