Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2024.050112
ISSN 2790-0932 Vol. 5, Issue 1: 103-111

Application and Practice of Data Mining Algorithms in Power Evaluation System

Dazhong Wang[a*], Yinghui Xu[b], Yongshuang Zhang[c], Peng Liu[d], Chuang Li[e], Baoliang Zhang[f]

China Electric Power Research Institute, Beijing 100192, China

[a]wangdazhong@epri.sgcc.com.cn, [b]xcfw@epri.sgcc.com.cn,
[c]szhb-zhangyongshuang@epri.sgcc.com.cn, [d]liupeng3@epri.sgcc.com.cn,
[e]lichuang@epri.sgcc.com.cn, [f]zhangbaoliang@epri.sgcc.com.cn

[*]corresponding author

Keywords: Power Review System; Data Mining Algorithms; Feature Engineering; Decision Tree Modeling

Abstract: As one of the most important functions of the power industry, power fault detection and anomaly identification are supported by power audit systems. In power audit systems, data mining algorithms are widely used to improve the identification accuracy as well as the stability of the system. This study examines the performance of the Decision Tree Algorithm, SVM Algorithm, and Bayesian Algorithm in determining the power fault in the power audit system. We used the dataset of power auditing system and divided it into training set and test set to build the classification model using Decision Tree Algorithm, SVM Algorithm and Bayesian Algorithm respectively and evaluated the test set to compare the three algorithms in terms of accuracy, precision and system stability. Experimental results show that the Bayesian Algorithm is outperformed by the Decision tree Algorithm and SVM Algorithm in the power audit system. The Decision tree Algorithm has better ability of feature processing and feature modeling. The Decision tree Algorithm can automatically select the best features for classification judgment.

1. Introduction

The power review system is the core of the power grid system that can be used for power data analysis and identification, fault detection and abnormal event identification and provide a basis for decision-making in operation and maintenance of power enterprises. However, precise detection of faults and abnormalities from power data is still an issue because existing power meters generate data that is diverse and complex. Power review systems could perform better with algorithms for data mining involved in the system, which helps find useful patterns and rules from massive data.

The aim of the paper is to compare the performance of the Decision Tree Algorithm, SVM Algorithm and Bayesian Algorithm on power review system and review their advantages and

Copyright: © 2024 by the authors. This is an Open Access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).

disadvantages in terms of accuracy, recall and system stability. This study has important significance. First, it can guide the choice of algorithm and optimization of a power review system by comparing the performance of different algorithms. Second, by understanding the advantages and disadvantages of different algorithms, it can further improve and innovate the design and application of power review systems. Third, the results of this study have practical significance for the power industry to improve the reliability of the power system, reduce the failure rate and improve operating efficiency.

This paper first introduces the background and related research on power review systems and discusses the application of data mining algorithms in this field. Next, the paper describes the research methodology in detail, including the collection and preprocessing of the dataset, and the modeling process of the decision tree algorithm, SVM algorithm and Bayesian algorithm. Then, the paper analyzes and discusses the experimental results, comparing the performance of different algorithms in terms of accuracy, recall and system stability. Finally, the paper summarizes the main findings of the study and proposes directions for further improvement and research in the future.

2. Related Work

Many studies have been conducted on the power review system. Wang Chao designed a cost evaluation system architecture with hierarchical interaction of "resources standards business data applications", explaining the implementation process of main functions such as economic evaluation, engineering budget review, electronic pre settlement, and data collection. He also introduced the technical architecture and permission settings of the system implementation, which can provide reference for the construction of digital evaluation systems and the digital transformation of cost management[1]. Yuan Yan proposed the intelligent evaluation management system of electric power project based on big data analysis. On the basis of constructing the logical structure of intelligent evaluation management system of electric power engineering, he formulated the evaluation standard of electric power engineering by using big data analysis, and carried out specific evaluation analysis [2]. Fang Ming analyzed the power project design review and technical and economic evaluation needs, and at the same time, described the power project design review and technical and economic evaluation information system, aiming to provide a reference for related research [3]. Zhang Rui proposed a visual evaluation method of power engineering materials based on BIM lightweight model. The proposed method used the BIM lightweight model to construct a 4D power engineering material supply plan model, simulate the execution of power engineering material supply plan, and then filter out the appropriate material visualization evaluation index [4]. Xiong Xiong analyzed the status quo and existing problems of power enterprise informatization project evaluation, focused on the feasibility study of power enterprise informatization project, the preliminary design stage results, from the technical program, the project cost of the two aspects of the comb and summarize the evaluation of key points, put forward for the review of informatization projects to improve the work of the proposal for the review of the electric power informatization project evaluation process and the future development of providing reference [5]. The aim of Fleischmann S was to examine the fundamental development of the concept of pseudocapacitance and its importance in electrochemical energy storage and to describe new materials whose electrochemical energy storage behavior can be described as pseudocapacitive [6]. Chen G comprehensively and deeply reviewed the research activities on using smart textiles to obtain energy from renewable energy sources in the human body and its surrounding environment [7]. Ufa R A described the impact of distributed generation on power losses, voltage levels, the possibility of maintaining power balance and participating in frequency regulation, and short-circuit currents in power systems [8]. Hatziargyriou N summarized the main results of the task force's work

based on the IEEE PES report and proposed an extended definition and classification of power system stability [9]. Zhao J discussed the advantages of DSE over static state estimation and the implementation differences between the two, including measurement configurations, modeling frameworks, and supporting software features [10]. These studies provide many reference methods for this work, which will build a power review system through data mining algorithms and compare the performance differences between different algorithms.

3. Method

3.1 Electricity Review System

The Power Review System is an application system based on data mining algorithms and artificial intelligence techniques designed to assess and review all aspects of the power industry in order to improve the reliability, efficiency and security of the power system. It collects, processes and analyzes power data to provide decision support and risk assessment to help the electric power industry achieve optimized operation and management. Thus, the power review system is able to collect and organize relevant data in real time from data interfaces accessing various parts of the power system, including power generation, transmission, distribution, and consumption [11]. At the same time, it preprocesses and cleans the data to ensure the accuracy and completeness of the data. Furthermore, it applies data mining algorithms and machine learning techniques to explore and analyze large amounts of power data. By exploring the correlations and patterns between data, the system is able to identify potential problems, abnormalities, and risks and provide early warnings and decision support. Commonly data mining techniques are used, including cluster analysis, classification algorithms, and association rule mining. In addition, by analyzing and modeling historical data, the power review system is able to provide risk assessment and prediction of the power system, and, by identifying potential risk factors and weak links, the system can take actions proactively to reduce potential failures and accidents and ensure the stable operation of the power system. Based on the results of data mining and analysis, the power review system provides decision support and optimization recommendations. Therefore, by analyzing and evaluating the key performance indicators and the key parameters of the power system to assist decision-makers in making a reasonable operation strategy, resource allocation, and scheduling plan to for the purpose to optimizing the economic operation of the power system.

3.2 Data Collection and Pre-processing

The power review system can collect real-time power data from various sensor devices, such as sensor data from generators, transmission lines, transformers, and other devices. These data can be collected in real-time through IoT technology and data interfaces. SCADA (Supervisory Control and Data Acquisition) system is a commonly used data acquisition and monitoring system in the power system, which can provide real-time data from various parts of the power system. The power review system can be interfaced with the SCADA system to obtain data from the relevant links [12]. Historical data can also be used for analysis and modeling. These historical data can be operational data, fault records, maintenance logs, etc. of the power system over a period of time in the past. Historical data can be acquired either by extracting it from a database or importing it from an archive file. During the data collection process, the raw data may have missing values, outliers, duplicates, etc., and data cleaning is required to ensure the quality and accuracy of the data. Standardization is a common method of data scaling that converts data to a standard normal distribution with a mean of 0 and a standard deviation of 1. Normalization removes the differences in magnitude between different features, making them have the same scale range, which helps the

convergence and performance of certain data mining algorithms:

$$x' = \frac{x - u}{\sigma} \quad (1)$$

x' is the normalized data, x is the original data, u is the mean of the original data, σ is the standard deviation of the original data. Table 1 shows the processed data presentation:

Table 1: Display of processed data

Timestamp	Generator Power	Transmission Line Voltage	Transformer Temperature
2022-01-01 09:00:00	75.12	432.56	68.24
2022-01-01 09:01:00	82.45	410.28	71.89
2022-01-01 09:02:00	79.68	425.13	75.36
2022-01-01 09:03:00	81.23	415.77	70.92
2022-01-01 09:04:00	77.95	430.09	72.81

Table 1 shows several fields for collecting real-time power data: timestamp, generator power, transmission line voltage, and transformer temperature.

3.3 Feature Engineering

Principal Component Analysis (PCA) is a commonly used feature extraction and selection method for converting high-dimensional feature data into low-dimensional features while retaining the most important information. In power review systems, PCA can be used to reduce the dimensionality of load data so as to extract the features that are most relevant to load variations, providing support for tasks such as load forecasting and anomaly detection [13]. The goal of PCA is to map the original feature data to a new feature space by linear transformation, which consists of a set of new features called principal components. These principal components are linear combinations of the original features that retain the maximum amount of information in the data by maximizing the variance. Supposing we have an m ×n feature matrix X, where m is the number of samples and n is the number of features, compute the covariance matrix C of the feature matrix X, which is defined as:

$$C = (X - \mu)(X - \mu)^T \quad (2)$$

Performing eigenvalue decomposition on covariance matrix C to obtain eigenvalue $\lambda_1, \lambda_2, ..., \lambda$□ and corresponding eigenvector $v_1, v_2, ...,$ v□. Multiplying the original eigenvalue matrix X with eigenvalue matrix P to obtain the dimensionality reduced eigenvalue matrix Y:

$$Y = X * P \quad (3)$$

By using PCA for feature extraction and selection, we can convert the original high-dimensional feature data into lower dimensional principal component features, thus realizing data dimensionality reduction and reducing redundant information while retaining the most important feature information.

3.4 Data Mining Algorithm Selection and Application

Decision Tree (DT) algorithm is a supervised learning algorithm based on tree structure, which

can be used for classification and regression tasks. The Decision Tree algorithm is widely used in power system analysis for tasks such as load classification, anomaly detection and energy consumption analysis. Decision Tree algorithms try to categorize power loads based on a number of features and attributes in the power system [14]. A decision tree model can be established with load data classified into different categories, such as high load, low load, peak load, etc. This helps power system operators to understand different load behavioural patterns, and to offer recommendations for power supply and dispatch. Apart from categorization function, when used in anomaly or fault detection, decision tree model training is used, which will identify abnormal load behavior that does not conform with normal operation patterns. This can pave the road to identify potential problems, like abnormal changes of load or defective equipment, in a timely manner, and invoke necessary remedies and control check measures. Decision tree models can be constructed to identify features and attributes which have significant impact upon energy consumption, such as weather conditions, seasonal variation, and equipment status [15]. This can be used to help power operators to understand energy usage patterns and demand trends and optimize energy strategies towards the reduction of energy costs and environmental burdens.

4.Results and Discussion

4.1 Experimental Design

The purpose of the experiments was to verify the accuracy and applicability of the decision tree algorithm in the load classification task, and to show how it compares to other classification algorithms. We want to collect power system data sets to include load data and related features sufficient to make the experiment have enough samples and diverse enough data to make the experiment reliable and well positioned for representativeness. According to the needs of the power monitoring system, the features that matched the load classification task were selected. In terms of the experimental environment, a server with a multi-core processor and sufficient memory capacity is configured, sufficiently large hard disk space is used to store the power system dataset and experimental results, Windows operating system is selected, Python is used for constructing the algorithmic model, and the DecisionTreeClassifier in Scikit-learn is used to support data processing, feature selection and model construction. In terms of experimental design, we used a control group experimental design. The experimental group used the decision tree algorithm to construct a load classification model and used the model to classify the test data for prediction. While the control group chose some representative classification algorithms, Support Vector Machine (SVM), Naive Bayes, and used these algorithms to classify the test data for prediction. Accuracy, recall, and system stability are selected as the evaluation indexes of the experiments, multiple experimental repetitions are performed, and the experimental results are calculated to minimize the influence of randomness on the results. Finally, the experimental results are recorded and counted for data analysis.

4.2 Experimental Results

Accuracy measures the correctness of the system in classifying or judging power data, and recall refers to the ratio of the number of positive samples correctly predicted by the power review system to the actual number of positive samples, and the combined consideration of accuracy and recall can provide an assessment of the overall performance of the power review system, Figure 1 and Figure 2 show the results of the comparison of the accuracy and recall, respectively:

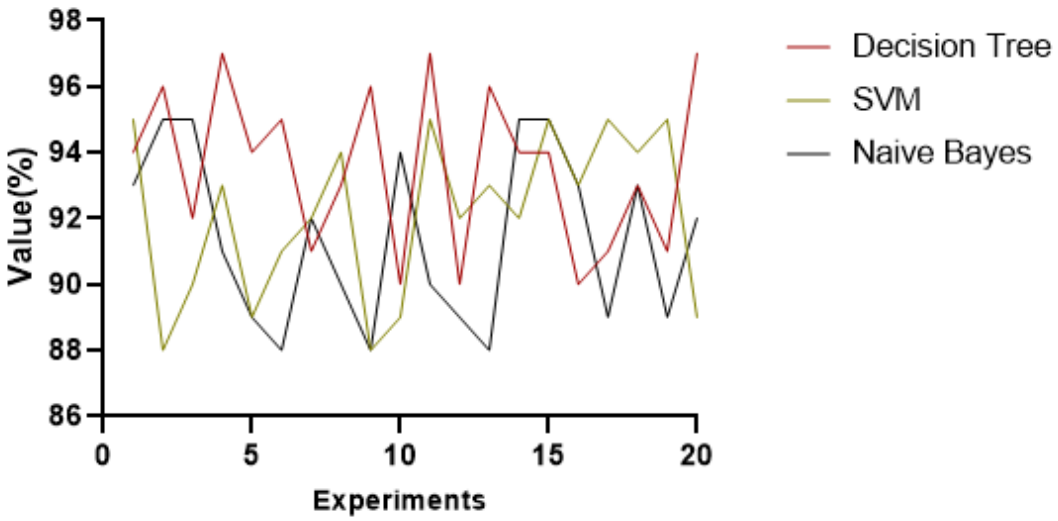

Figure 1: Accuracy

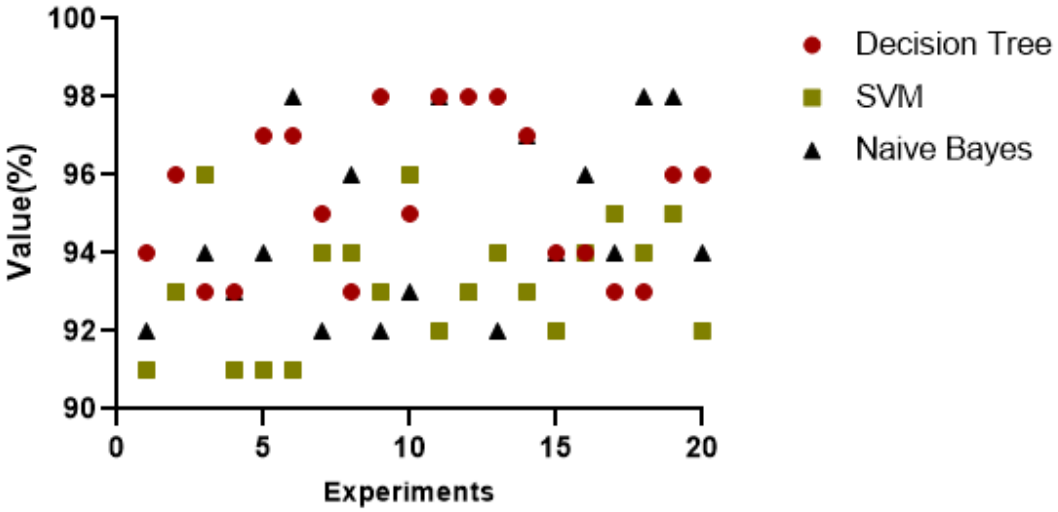

Figure 2: Recall rate

In testing, the decision tree algorithm in this paper performs between 90%-97% accuracy and 93%-98% recall, the support vector machine algorithm performs between 88%-95% accuracy and 91%-96% recall, and the Bayesian algorithm performs between 87%-95% accuracy and 92%-98% recall. Based on the comparison results, it can be seen that the decision tree algorithm performs best in terms of accuracy and recall. This is due to the fact that the decision tree algorithm takes into account the interaction and nonlinear relationship between features when constructing the decision tree model. This enables the decision tree algorithm to better capture the complex patterns and regularities in the power data and improve the classification accuracy and positive sample identification.

Figure 3 shows the results of system stability comparison:

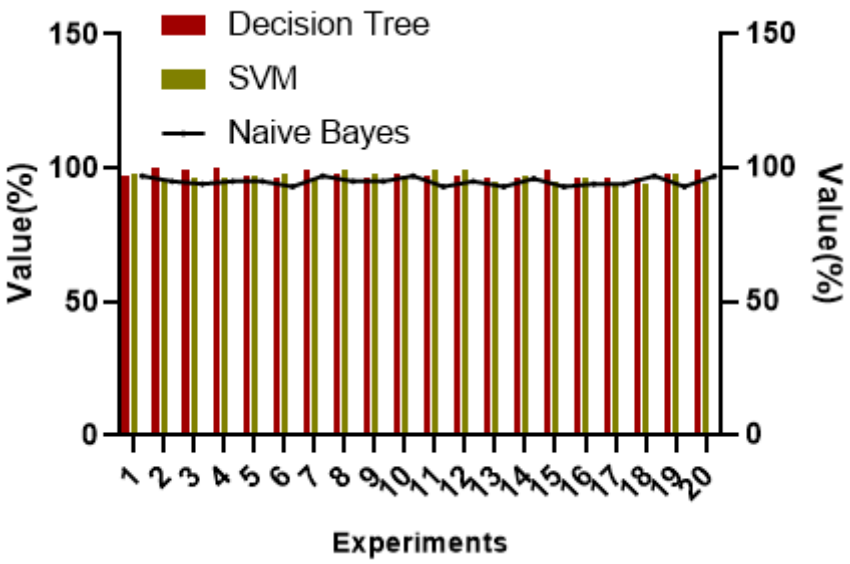

Figure 3: System Stability

The integrity tests for comprehensive stability showed that the highest stability belonged to the decision tree algorithm proposed in this paper with 100% for the maximum, 96% for the minimum, and 99.75% in average stability. This means that the decision tree algorithm has produced consistent and reliable results in many runs and tests, showing very high stability. It is important to guarantee stability for the power review system, so stability means how consistent and reliable the system is in different situations, which decreases the misjudgment and uncertainty. In contrast, the SVM algorithm has a lower stability than the decision tree algorithm at the highest 99%, lowest 94% and 97.21% average. Although it is not as stable as the decision tree algorithm, the SVM algorithm still has relatively high stability, which will be able to provide stable results in different test situations. The Bayesian algorithm that uses all previous data with different levels of significance has 97% for the maximum, 93% for the minimum, and 95.72% for average stability. Although the Bayesian algorithm was lower among the three algorithms in their respective stability, overall it is still stable. Comparing the other two compared algorithms, the Bayesian algorithm has stable behavior meaning that it is relatively consistent to provide consistent results as designed for the power review system, regardless of the test situation.

4.3 Summary

Decision tree algorithm has certain advantages in feature processing and modeling. It is able to construct classification rules by splitting features to classify and judge the power data, because it can automatically pick out the best feature for splitting, make the decision according to the importance of feature, then the model will fit the data better. On the contrary, SVM and Bayesian algorithms may be weak in terms of feature processing and modeling, as they cannot fully utilize interactions and non-linear relationships between features.

5. Conclusion

Data mining algorithm in power-usage review systems has been significantly applied and experimented with conceptually. After evaluating indicators like precision, recall, and system stability, the decision tree algorithm reformed outstandingly when compared with SVM and

Bayesian algorithms. The success of the application of decision tree algorithm in the power-usage review system is mainly attributed to the processing of features, abilities of modeling, interpretability, adaptability for data distribution and the fitting of the model, and stability of the system. Evidently, the increase in precision and recall revealed that the decision tree algorithm can accurately identify the states of power faults and anomalies and improve the accuracy of system decision. Also, the interpretability of the decision tree algorithm can make the decision based on the model more transparent and credible, which in turn improves the reliability of the system. Finally, due to the improved stability of the algorithm, it guarantees the consistency and reliability of the system results in different tests.

References

[1] Wang Chao, Zhang Jiyuan, Zhou Ping, Wang Yun, Guo Xiaofan, Wu Huaqiang. Research on the Construction of a Whole Process Digital Electric Power Cost Evaluation System [J]. Building Economics, 2023, 44 (1): 91-98.

[2] Yuan Yan, Zhang Chun, Peng Peng, Guo Junfeng, Wang Weifeng. Design of Intelligent Evaluation and Management System for Power Engineering Based on Big Data Analysis [J]. Wireless Internet Technology, 2024, 21 (3): 12-14.

[3] Fang Ming, Zhang Shuangping, Lu Qiuyun, Zhu Deyan, Ye Jian. A Brief Discussion on the Information System for Design Review and Technical Economic Evaluation of Power Engineering [J]. China Management Informatization, 2019, 22 (12): 56-57.

[4] Zhang Rui, Ji Jianren, Zhu Jixiang, Li Jiashuan, Chen Yi, Wei Jun. Visual evaluation method for power engineering materials based on BIM lightweight model [J]. Microcomputer Application, 2023, 39 (3): 28-31.

[5] Xiong Xiong, Zhao Dan, Meng Anning, Hou Jinxiu, Zheng Jiapeng. Key Points and Suggestions for Evaluation of Informationization Projects in Electric Power Enterprises [J]. China Management Informatization, 2023, 26 (16): 112-115.

[6] Fleischmann S, Mitchell J B, Wang R, et al. Pseudocapacitance: from fundamental understanding to high power energy storage materials[J]. Chemical Reviews, 2020, 120(14): 6738-6782.

[7] Chen G, Li Y, Bick M, et al. Smart textiles for electricity generation[J]

[8] Ufa R A, Malkova Y Y, Rudnik V E, et al. A review on distributed generation impacts on electric power system[J]. International journal of hydrogen energy, 2022, 47(47): 20347-20361.

[9] Hatziargyriou N, Milanovic J, Rahmann C, et al. Definition and classification of power system stability–revisited & extended[J]. IEEE Transactions on Power Systems, 2020, 36(4): 3271-3281.

[10] Zhao J, Netto M, Huang Z, et al. Roles of dynamic state estimation in power system modeling, monitoring and operation[J]. IEEE Transactions on Power Systems, 2020, 36(3): 2462-2472.

[11] Ozcanli A K, Yaprakdal F, Baysal M. Deep learning methods and applications for electrical power systems: A comprehensive review[J]. International Journal of Energy Research, 2020, 44(9): 7136-7157.

[12] Sepulveda N A, Jenkins J D, Edington A, et al. The design space for long-duration energy storage in decarbonized power systems[J]. Nature Energy, 2021, 6(5): 506-516.

[13] Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, et al. Potential value and impact of data mining and machine learning in clinical diagnostics[J]. Critical reviews in clinical laboratory sciences, 2021, 58(4): 275-296.

[14]Barzkar A, Ghassemi M. Components of electrical power systems in more and all-electric aircraft: A review[J]. IEEE Transactions on Transportation Electrification, 2022, 8(4): 4037-4053.

[15]Ramoji S K, Saikia L C. Maiden application of fuzzy-2DOFTID controller in unified voltage-frequency control of power system[J]. IETE Journal of Research, 2023, 69(7): 4738-4759.