Proceedings of the 2nd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/53/20240135

Research on BP neural network model-based data mining technique in KOL identification

Jiarui Hou[1], Tianyu Li[1,2], Yahui Wu[1], Pei Wang[1]

[1]School of Management Science and Engineering, Baoding University of Technology, Baoding, China

[2]757314166@qq.com

Abstract. Accompanied by the development of artificial intelligence industry, to play the data value of web crawler technology has been the focus of research in the field of computer network and data science, and the data mining technology based on artificial neural network intelligent algorithm is widely used. In view of this, this paper takes the BP neural network data mining technology, which has excellent nonlinear mapping capability, parallel processing capability and fault tolerance and is widely used, as the basis, and integrates the methods and ideas of data mining into the mining of data laws in the field of KOL identification of key opinion leaders, with a view to finding valuable intrinsic laws and relationships between the mining of web crawler technology and the identification of KOL features. The research content of this paper mainly includes two aspects of research work, the design of high-performance data mining technology and the actual work in the field of KOL recognition. On the one hand, this paper comprehensively describes the basic theory and methods of data mining, and focuses on the in-depth analysis and elaboration of BP neural network-based data mining technology on the basis of understanding and analyzing a variety of data mining technologies. On the other hand, this paper aims to solve the problem of bias in the prediction of traditional KOL model and design the experimental method of KOL recognition by BP neural network algorithm.

Keywords: BP neural network, web crawler, key opinion leader, data value.

1. Introduction

In recent years, with the rapid development of artificial intelligence and technology, various scientific research fields have more diverse methodologies. Among them, data mining based on big data has become one of the fields with great potential for future development. Data mining, also known as knowledge discovery in databases, reveals the hidden patterns and relationships in the raw data by intelligently processing and analyzing them, so as to achieve effective classification, correlation, and prediction of the target problem, and ultimately enhance the value of the data [1].

Big data-driven data mining techniques have shown great vitality in today's information age, and the process mainly consists of three stages: data preparation, data mining, and interpretation and evaluation of results. However, it is often challenging to utilize basic databases such as scientific books and research papers, which are similar to low-quality “ores” rather than high-quality “gold mines”. How to effectively develop these low-value, high-capacity data ores and explore their potential value is one of the major problems and challenges in the field of data mining [2].

DOI: 10.54254/2753-8818/53/20240135

Currently, data mining in the basic sciences faces several major problems:

First, database quality issues. The databases used for data mining usually integrate various data sources, so there may be uncertainties in their data quality, including data accuracy, measurement errors, collection errors, and duplicate data [3];

Second, specialized database attributes. Data mining for a specific target problem requires specialized knowledge in related fields, covering data collection, representation, and subsequent processing;

Third, the complexity of related variables. When dealing with specific target problems, data mining usually involves numerous complex related variables, and the relationship between these variables and the target problem is generally not linear;

Fourth, immaturity of data mining theories and methods. Despite some progress, there is still a lack of universal or standardized algorithmic models that can effectively deal with realistic challenges such as large data volume, noise interference, variable complexity and specialized target characteristics [4].

In recent years, significant progress has indeed been made in the field of data mining, especially in terms of algorithmic diversification and in-depth theoretical research. The application of BP neural network algorithms in data mining has indeed demonstrated its advantages in dealing with complex, nonlinear problems. In the early data mining research, researchers usually used the combination of database and symbolic machine learning. However, the universality of these methods is limited when facing the problems of many data variables, complex nonlinear relationships, and noisy data interference in real cases. In recent years, scholars at home and abroad have made continuous efforts to propose many new data mining algorithms and construct diverse theoretical systems. The data mining methods with more applications at present include traditional statistical methods, rough set theory methods, decision tree theory methods, and artificial neural network theory methods. Artificial neural network methods, especially BP (Back Propagation) neural network algorithm, show obvious advantages and practicality in data mining. BP neural network simulates the signal processing and processing ability of the biological brain, and it has good association, storage and self-learning ability, so it is more suitable for dealing with nonlinear and complex problems. BP neural network is able to find out intrinsic patterns through learning the database By learning the information structure in the database, BP neural network is able to find out the inner law and analyze and predict, without the need to deeply understand the internal complex mechanism characteristics of the information system. This makes BP neural networks widely used and rapidly promoted in the fields of industrial engineering, discipline intersection and complex problem optimization. Overall, with the continuous development of data mining algorithms and theories, BP neural network, as a powerful tool, shows its unique advantages and potentials in dealing with complex data analysis and prediction in reality.

Based on previous research, this paper focuses on using web crawler technology to obtain e-commerce operation data and develops a BP neural network with excellent performance through the method of fusing the advantages of multiple algorithms. It aims to introduce data mining theories and techniques into the field of KOL and establish an analytical model for law mining of KOL identification data to reveal valuable potential laws behind the data. This kind of law mining relies on a large amount of experimental data and has higher reliability than the mechanism prediction model derived from the single variable control method. More importantly, the rational use of the mined hidden laws can not only effectively correlate and analyze various potential influencing factors and provide new ideas and directions for experimental research, but also effectively predict the characteristics of unknown samples. Through the data mining method, it not only significantly reduces the cost and workload of KOL identification, but also expands the theory and method of KOL data law analysis, and realizes the efficient use of related data. Therefore, this study has important academic value and practical significance for the in-depth exploration of data laws obtained by web crawlers.

DOI: 10.54254/2753-8818/53/20240135

2. Experimental design

2.1. Data sources

In this paper, according to the target information of the enterprise needs to automatically capture the platform to create content dissemination data, the idea is to use web crawler technology to collect data. After building BP neural network through data cleaning, the general input layer information are: the amount of likes, comments, retweets, author's mood, account level, creation quality, heat persistence, these indicators are used to determine whether the published user information can be highlighted in a large number of information. The data mining stage is the core part of the research, and its core purpose is to explore patterns of information that are useful or valuable to the organization from the established data set. In this process, first of all, we have to construct a mining model, specifically, we have to choose some specific variable factors to realize the mining target task, transform the data set into an analytical model, and then carry out data mining based on some mining algorithms on the analytical model. Due to the restriction of "anti-crawling mechanism" and "China's law on crawlers", this paper cleansed the text data of 50 key opinion leaders recognized in the beauty industry and 50 users with general publicity effect for the training of BP neural network model, the indicators of KOLs, the judgment basis and sample descriptive statistics. judgment basis and sample descriptive statistics are shown in Table 1 below.

Table 1. Description of general user data samples for training BP neural network models

	Mean	Std	Min	Max
Volume of comments	0	0	0	0
Forwarding volume	0	0	0	0
Likes	0	0	0	0
Author emotion	1	0	1	1
Level of account	1	0	1	1
Creative quality	0	0	0	0
Persistence of heat	0	0	0	0

Table 2. Description of sample KOL user data for training BP neural network models

	Mean	Std	Min	Max
Volume of comments	1	0	1	1
Forwarding volume	0.92	0.27	0	1
Likes	1	0	1	1
Author emotion	0.98	0.14	0	1
Level of account	1	0	1	1
Creative quality	0.58	0.49	0	1
Persistence of heat	0.94	0.23	0	1

As shown in Tables 1 and 2, KOL users reflect superiority over general users in terms of comments, retweets, likes, author's sentiment, quality of creation, and sustained heat of content.

2.2. BP Neural Network Modeling Algorithm

BP Neural Networks, which originated from the M-P model in the 1940s, is an important discovery in the field of Artificial Intelligence and a landmark achievement of human beings in the field of computational tools. The working principle of artificial neural networks is more similar to the working process of the real biological nervous system (human brain), which gives it the ability to process and handle information similar to that of the human brain; it accomplishes the complex non-linear target tasks in reality through continuous iterative digitized operations with the help of computer software [5].

DOI: 10.54254/2753-8818/53/20240135

Neural network is actually a large number of bionic artificial neurons, which in essence is a simplification and abstraction of the characteristics of real biological neurons. Artificial neurons are the basic unit of neural network integrating information storage, conversion and processing, which is characterized by parallel processing in time and distributed storage in space. The parallel processing ability makes the neural network has a faster running speed, and shows the diversity of functions; distributed storage ability makes the neural network has an excellent fault tolerance performance, that is, when the input signal of some neurons in the network is fuzzy or deformed, the effective real information can still be obtained in other neurons and give the approximate solution, which is similar to the ability of the human brain to recall the information from some information fragments to recover to close to the information of the whole picture. This is similar to the ability of the human brain to recover from certain information fragments to the full picture of information when recalling. Thus, these features give artificial neural networks the ability to be self-organizing, self-adaptive, distributed memory (storage), and parallel processing of complex information.

In order to match the appropriate activation function to represent the decision boundary of KOL recognition and to fit any smooth mapping with any accuracy, in this paper, the number of hidden layers is set to 2, and each layer is set to 6 neurons. Based on the following equation (1), this paper passes the variables such as the number of likes, comments, retweets, author's sentiment, account level, creation quality, and heat persistence into the input layer of the BP neural network, and solves the variables of the output layer inversely based on the following equation (2), and the activation function is shown in the following equation (3) [6].

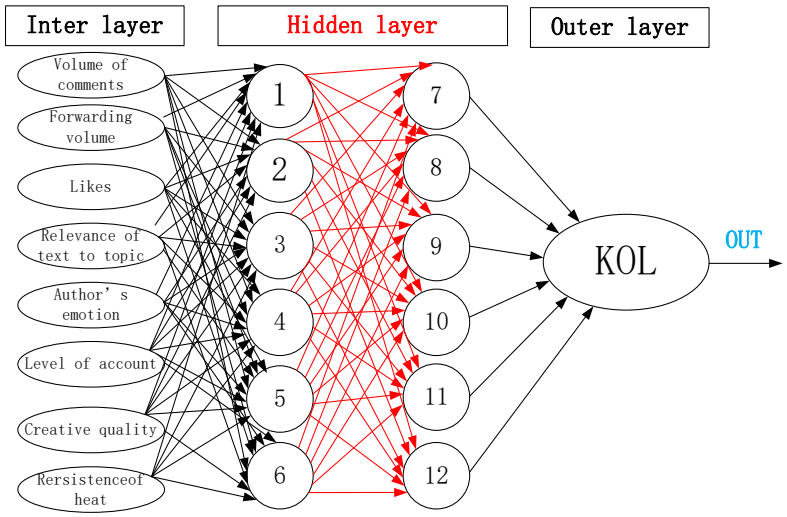

Figure 1. Training parameter settings for the artificial neural network model

$$\theta = \sum_{i=1}^{m} W_i x_i + b \tag{1}$$

$$y = f(\theta) = f\left(\sum_{i=1}^{m} w_i x_i + b \right) \tag{2}$$

$$y = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b \tag{3}$$

3. Experimental analysis results

In this paper, we use the rule of thumb to set the number of hidden layers to 2 and the number of neurons in each layer to 6. The parameters and bias after training the BP neural network using the training set are shown in Figure 3 below.

DOI: 10.54254/2753-8818/53/20240135

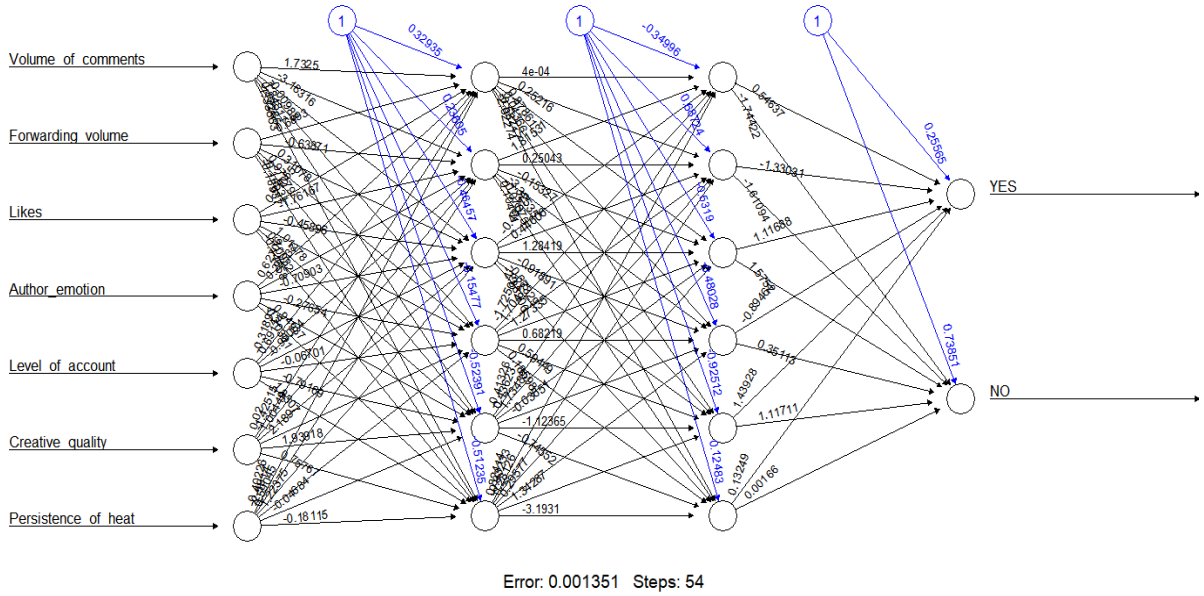

Figure 2. Visualization process of BP neural network training results

As shown in Figure 2 above, the prediction results of the BP neural network show a very high superiority with an error of only 0.001351.

As shown in Figure 3 below, the prediction accuracy by the artificial neural network model on the 10% prediction set is approximated to be 100%.

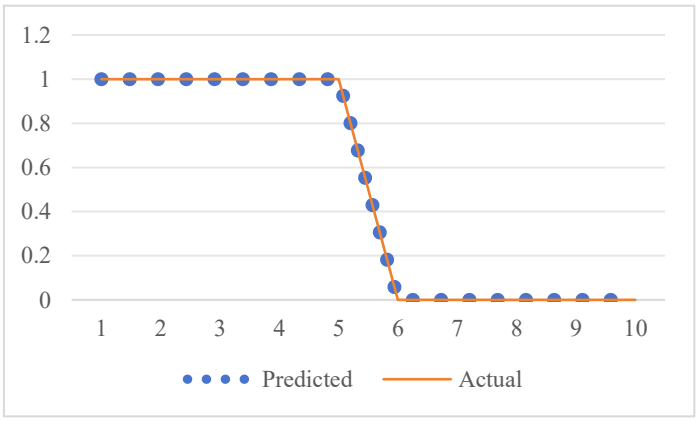

Figure 3. Predictability of BP neural network models

In this paper, the process of identifying Key Opinion Leaders (KOLs) combines web crawling techniques and BP neural network algorithms to ensure accuracy and real-time performance. On the one hand, web crawling techniques are used to collect large-scale user data from various social media and online platforms. These data include not only users' interactive behaviors such as comments and likes, but also the content they post and their text features. After data cleaning and preprocessing to remove noise and outliers, the data preparation makes the subsequent analysis more effective and reliable. In the future, BP neural network algorithms can be utilized to extract key features from the prepared data. This includes the extraction of textual features, such as keywords and sentiment analysis, and user behavioral features, such as interaction frequency and patterns. By building and training neural network models, the influence of each user can be evaluated and ranked to identify the most influential opinion leaders. These leaders are not just a reflection of how many followers they

DOI: 10.54254/2753-8818/53/20240135

have, but also their deep influence and trust in a specific domain or topic. In addition, it is necessary to ensure the timeliness and accuracy of the identification results by continuously optimizing the model and updating the data [7].

4. Conclusion

Based on the experimental results and data analysis described in this paper, the BP neural network structure using the rule-of-thumb setup (the number of hidden layers is 2 and the number of neurons per layer is 6) shows significant superiority on the training set. The prediction results show that the error of this BP neural network is only 0.001351 on the test set, while the prediction accuracy is close to 100% on the 10% prediction set. These results indicate that the employed neural network model is highly effective and accurate for the problem addressed, providing a strong reference and guidance for future applications of similar tasks.

References

[1] Yingli Wu and Xin Li and Qingquan Liu and G. Tong The Analysis of Credit Risks in Agricultural Supply Chain Finance Assessment Model Based on Genetic Algorithm and Backpropagation Neural Network Computational Economics (2021):60.

[2] Song, Yaoling, Yage Jing, and Xuan Qin. “BP neural network-based early warning model for financial risk of internet financial companies.” *Cogent Economics & Finance* 11.1 (2023): 2210362.

[3] Muder Almiani and A. Ghazleh and Y. Jararweh and A. Razaque DDoS detection in 5G-enabled IoT networks using deep Kalman backpropagation neural network International Journal of Machine Learning and Cybernetics (2021):12.

[4] Saipraneeth Gouravaraju and Jyotindra Narayan and R. Sauer and S. Gautam A Bayesian regularization-backpropagation neural network model for peeling computations The journal of adhesion (2020):99.

[5] Lin Wang and Binrong Wu and Qing Zhu and Yurong Zeng Forecasting Monthly Tourism Demand Using Enhanced Backpropagation Neural Network Neural Processing Letters (2020):52.

[6] Shangyu Zhao and Guoying Chen and Mingya Hua and C. Zong An identification algorithm of driver steering characteristics based on backpropagation neural network Proceedings of the Institution of mechanical engineers. Part D, journal of automobile engineering (2019):233.

[7] Muhammad Yuslan Abu Bakar and Adiwijaya and S. A. Faraby Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation)Using Information Gain and Backpropagation Neural Network International Conference on Asian Language Processing (2018).