Research and Application of Investment Risk Identification Based on Probability-Deep Neural Network

Yuheng Sha[1,a] *, Qian Ma[2,b], Chao Xu[3,c], Xue Tan[4,d], Jun Yan[5,e], Yuqian Zhang[6,f]

*[a] Corresponding author: Yuhengsha@126.com, [b]e-mail: 1454846725@qq.com,
[c]e-mail: 353287454@qq.com, [d]e-mail: 1215030617@qq.com, [e]e-mail: 1770269953@qq.com,
[f]e-mail: 13652089007@163.com

[1]State Grid Corporation of China Beijing, China
[2]State Grid Jiangsu Electric Power Co., Ltd. Nanjing, China
[3]State Grid Jiangsu Electric Power Co., Ltd. Nanjing, China
[4]State Grid Energy Research Institute Co., Ltd. Beijing, China
[5]Tianjin Tianda Qiushi Power New Technology Co., Ltd. Tianjin, China
[6]Tianjin Tianda Qiushi Power New Technology Co., Ltd. Tianjin, China

Abstract—As the scale of power grid construction continues to expand, the scale of investment is also increasing, and investment risk factors are becoming more and more complex. To ensure the efficiency of power grid investment, it is necessary to identify power grid investment risks and formulate relevant preventive measures. Firstly, it uses natural language processing (NLP) and term frequency- inverse document frequency (TF-IDF) method to preprocess the text. Secondly, probabilistic neural network (PNN) is used to divide text materials into five kinds. Finally, back-propagation (BP) is used to evaluate the risk level of each text, achieving the risk category and risk level of the project investment. Furthermore, on this basis, the investment risk decision matrix is drawn and the investment risk response strategies are put forward, and an effective way for the practice of investment risk management in power grid projects are provided.

Keywords-Project risk identification; Text classification; Neural network; Risk decision matrix; Grid investment

1. Introduction

How to scientifically and effectively identify, analyze and deal with investment risks is an important topic in the practice of grid investment risk management today. With the continuous advancement of power market reform, the impact of internal and external environments on power grid investment is becoming more and more complex, and the investment risks faced by power grid companies are also increasing. Therefore, it is necessary to identify investment risks before making investment decisions to help internal managers make scientific, reasonable, economical and effective decisions [1].

From the perspective of domestic and foreign research practice, the research on risk mainly starts from risk factors.MA and YANG deeply study the status of construction projects during the construction process, and identifies the risk factors according to the risk identification

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334474

process [2]. LI and LIU construct a risk matrix, puts forward the principles of risk assessment, evaluates the risk level according to the principles, and ranks the risk factors [3]. The above research provides a reference for the identification of investment risks, but most of the research relies heavily on historical data and expert experience, which is highly subjective and cumbersome in practical application.

At present, some studies have adopted certain methods to identify and evaluate grid investment risks. WANG and MA use the Monte Carlo method to construct a risk investment model for power grid investment concerning influencing factors such as electricity sales, electricity purchase price, transmission and distribution price, and line loss rate [4]. ZHANG and AN use the probability distribution model to realize the risk assessment of distribution network operation. However, the methods used in the above literature are not efficient and accurate and are not conducive to decision-makers to assess and deal with risks in a timely and effective manner [5]. Intelligent algorithms such as artificial neural networks provide new solutions for investment risk identification.

In order to solve the problem of grid investment risk identification, this paper will be based on the Python language. First, the NLP (Natural Language Processing) toolbox is called to process unstructured data such as cleaning, segmentation, and moving stop words in turn, and TF-IDF (Term Frequency- Inverse Document Frequency) is used for word frequency decomposition. Then, the text is divided into 5 categories according to the word frequency by PNN (Probabilistic Neural Network), and the key risk factors affecting investment are effectively identified. Through the application of the research results, the relevant requirements of power grid investment risk management can be connected with the specific measures of investment risk control, and the efficiency of investment decision-making can be improved.

2. Establishment of risk identification model

2.1 Text Preprocessing

To conform to the input form of the neural network, first, preprocess the classified text to filter out the effective feature itemsets of the text used to extract feature vectors. In this paper, NLP is used to complete the accurate word segmentation of the text, and then TF-IDF is used to extract the text feature vector [6, 7].

As a branch of artificial intelligence, NLP is mainly used to solve the problem of human-computer interaction using natural language. Word segmentation is the primary task of natural language processing technology. Use Python's Jieba module to implement Chinese word segmentation for risky text. Jieba is a free and open-source Chinese word segmentation module specially developed for Python. It is mainly based on the directed acyclic graph, dynamic programming to find the maximum path, Viterbi and other algorithms to achieve word segmentation and part-of-speech tagging.

To analyze the importance of a feature item t to a document d in a text set D, it is usually described by TF-IDF. When the number of occurrences of the feature item t in a certain document increase, its importance will increase, and when the number of occurrences in the entire text set increases, its importance will decrease. To judge the importance of a word or phrase in a document depends not only on the number of times it appears in the document but also the number of times it appears in other documents in the text set where the document is located. Only when the number of occurrences is small, the word can be judged to be of high importance and can be used as a key feature with a good degree of discrimination to classify the text.

2.2 Establishment of Risk Classification Model Based on PNN

PNN were first proposed by Dr. D.F. Specht in 1989 and are often used for pattern classification [8]. The PNN network generally consists of four layers, as shown in Fig. 1.

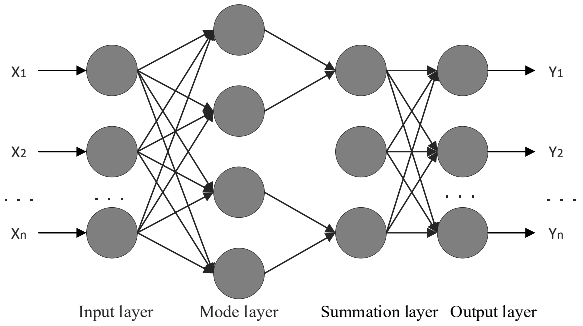

Figure 1. PNN network structure diagram

Input layer: The sample data enter the PNN network through the input layer, so the number of neurons in the input layer is the same as the number of variables in the sample data. It is used to connect external sample data, without any processing, and directly send the data to the mode layer.

Mode layer: The mode layer is the core layer of the entire PNN network. It compares the feature vector obtained by the input layer with the samples, collects similar samples, and calculates the similarity of the samples as the mode output by the following formula. Among them, the number of neurons is the number of samples multiplied by the number of classification categories.

$$f\left(x,W_{\mathrm{i}}\right)=\exp\left(\frac{\left(X-W_{\mathrm{i}}\right)^{T}\left(X-W_{\mathrm{i}}\right)}{2\sigma^{2}}\right) \tag{1}$$

In the formula, σ is the smoothing factor, which is a key parameter that determines the accuracy of the classification in the PNN network.

Summation layer: The summation layer is to add the output of the mode layer to obtain a whole, and obtain the estimated probability density of this type by the following formula according to the summation result. Among them, the neurons in the summation layer are only connected to the neurons in the pattern layer of the same category, and the number of neurons in the summation layer is the number of classification categories.

$$f_{A}\left(x\right)=\frac{1}{\left(2\pi\right)^{P/2}\sigma^{P}}\frac{1}{m}\sum_{i=1}^{m}\exp\left(-\frac{\left(X-X_{m}\right)^{T}\left(X-X_{m}\right)}{2\sigma^{2}}\right) \tag{2}$$

Output layer: The output layer determines the category of the final classification. It changes the confidence of the classification result by setting a threshold, and the output is the category with the highest probability density.

PNN has good generalization performance. When the number of samples changes, only the number of neurons in the mode layer needs to be changed, and when the number of sample categories changes, only the number of samples in the mode layer and output layer needs to be changed.

2.3 Determination of Risk Degree Based on BP(Back-Propagation) Neural Network

Artificial neural network is a model based on the human brain [9].

$$\vec{y} = f_{\text{network}}\left(\vec{x}\right) \tag{3}$$

In the input layer, because the model is nonlinear data, sigmoid is selected as the activation function.

$$f_{\text{u}} = \left[1 + \exp\left(-u\right)\right]^{-1} \tag{4}$$

Let w be the random weight between the input layer and the hidden layer and between the hidden layer and the output layer, its vector is $\vec{w}$, and the calculation formula of the hidden layer node O is shown in the following formula.

$$O = \text{sigmoid}\left(\vec{w}^{T} \bullet \vec{x}\right) \tag{5}$$

Similarly, the calculation formula of the output layer node P is shown in the following formula.

$$P = \text{sigmoid}\left(\vec{w}^{T} \bullet \vec{O}\right) \tag{6}$$

After the output result is obtained, the node output error δ is calculated according to the following formula.

$$\delta = P\left(1-P\right)\left(T-P\right) \tag{7}$$

Where T is the target value of the output node.

For the hidden layer node, its output error is shown in the following formula.

$$\delta_{\text{i}} = a_{\text{i}}\left(1-a_{\text{i}}\right) \sum_{\text{k}\in\text{outputs}} w_{\text{ki}}\delta_{\text{k}} \tag{8}$$

Let the hidden layer node be i, δ_{i} is the error term, a_{i} is the output value, w_{ki} is the connection weight between nodes, δ_{k} is the error term between nodes, and the weights on each link are updated as shown in the following formula.

$$w_{\text{ji}} \leftarrow w_{\text{ji}} + \eta\delta_{\text{j}}x_{\text{ji}} \tag{9}$$

In the formula, w_{ji} is the weight between nodes; η is the learning rate constant; δ_j is the error term of the node j; x_{ji} is the input between the nodes.

According to the method of determining the risk loss range, the power grid investment risk level recognition model based on BP neural network is set to three layers: input layer, hidden layer, and output layer. The input layer is five risk classifications and the loss range of each risk classification.

The flowchart of the risk identification model is shown in Fig. 2.

- Input text, the NLP toolbox is used to sequentially clean, segment, and move unstructured data such as stop words.
- Use TF-IDF to extract text feature vectors to judge the importance of words in the text.
- PNN is used to classify risks, and BP is used to determine the degree of risk.
- Build a risk decision matrix. And take corresponding risk treatment measures according to the decision matrix.

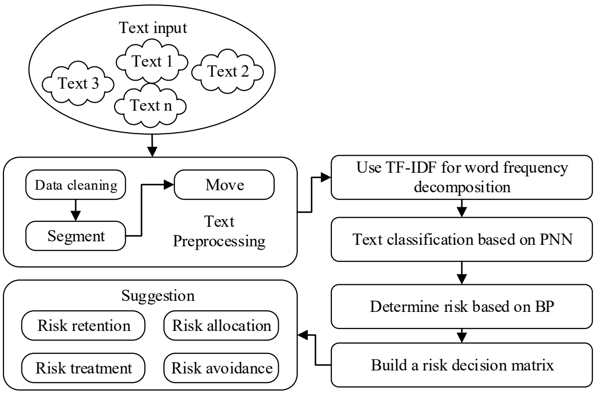

Figure 2. Flowchart of the identification model

3. Risk Decision Matrix

The risk matrix combines quantitative and qualitative methods, which can represent risks more intuitively. On this basis, this paper constructs the grid investment risk decision-making matrix from the two dimensions of grid investment risk classification and risk degree, as shown in Fig. 3.

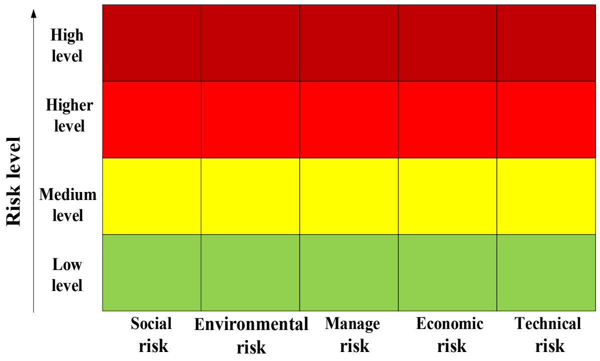

Figure 3. Grid investment risk decision matrix diagram

The grid investment risk response suggestions corresponding to the risk decision matrix are as follows:

- Green area

Take risk retention measures, the risk disposal method of the project investment manager prepares unforeseen expenses by himself to bear the risk losses.

- Yellow area

Take risk allocation measures.

- Orange area

Adopt risk control strategies, and adopt risk disposal techniques in order to minimize investment risks so as to reduce the extent of losses.

- Red area

Adopt risk avoidance strategy, when the investment will cause great losses, take the initiative to give up or terminate the investment or part of the investment to avoid losses.

4. Case analysis

This paper randomly selects 50 documents from the investment risk database of power grid companies. First, the documents are preprocessed, and the characteristics of risk sources are analyzed. The main risk sources are social, environmental, management, economic and technical risks. The classification results as shown in Table 1.

Table1 Grid investment risk classification

Risk source	**Risk source characteristics**			
Social risk	S1:Planning	S2:Tendering and bidding	S3:Expropriation of land	S4:Relocate
Environmental risk	E1:Policy	E2:Law	E3:GDP	E4:Urbanization level
Manage risk	M1:Organizational	M2:Operation	M3:Contract	M4:Security

	management	management	management	management
Financial risk	F1:Capital cost	F2:Capital needs	F3:Capital chain	F4:Capital turnover
Technical risk	T1:Technical difficulty	T2:Technology achievement maturity	T3:Technical implementation basis	T4:Use of new technology

Taking text 5 as an example for analysis, the classification result is shown in Fig. 4. Since the risk source characteristics of technical risks in text 5 account for a high proportion, text 5 is classified as technical risk.

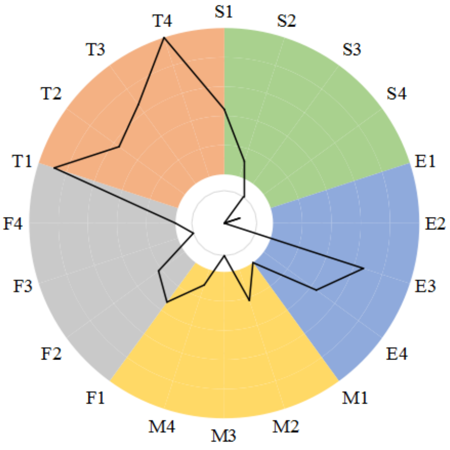

Figure 4. Risk Classification Results for Text 5

By adopting the method of PNN, the text is classified, the first 40 groups in the sample data are trained, and the last 10 groups of samples are used for verification. The error between the sample data and the predicted data is calculated as the basis for judging the training effect of the model. As shown in Fig. 5, the error after classification training is 97.5%, which can accurately reflect the classification results.

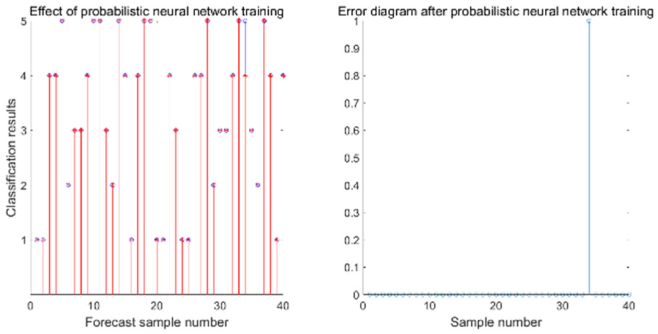

Figure 5. Text classification map based on probabilistic neural network

Use Matlab software to train the first 40 text scores, observe whether the convergence effect after 100,000 times of learning meets the requirements, and then verify according to the other

10 risk files after BP neural network training. The training error curve is shown in Fig. 6, the error of the test results is within 5%, and the network is reliable.

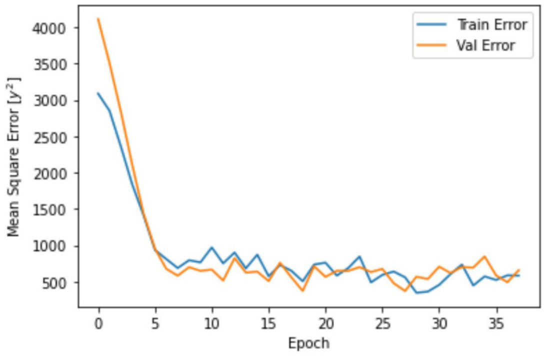

Figure 6. Error curve

After risk classification and risk degree analysis, the result of the risk decision matrix is shown in Fig. 7. The texts mainly focus on low risk and medium risk, and there is one text focused on medium risk and higher risk, which is environmental risk; there are two texts on management risk and economic risk, which are low risk and high risk respectively. 50% of risk texts focus on technology risk. According to the analysis, we should pay more attention to environmental risks and technical risks in the subsequent investment decision-making process, and take effective measures promptly. For example, when making power grid investment decisions, it is necessary to fully investigate the region's history and current policies and laws, and to conduct a detailed understanding and analysis of the local GDP, urbanization level and future development trends. Focus on technology research and development and innovation, as well as the prediction of technical difficulty and the maturity of technological achievements in the implementation stage, to ensure that the current technology implementation basis and power grid construction needs achieve the best degree of use and matching. Real-time tracking is carried out in the stage of using new technologies, the responsible subjects are clarified and refined， investment risks are identified and effective risk response plans are selected, to minimize investment risks.

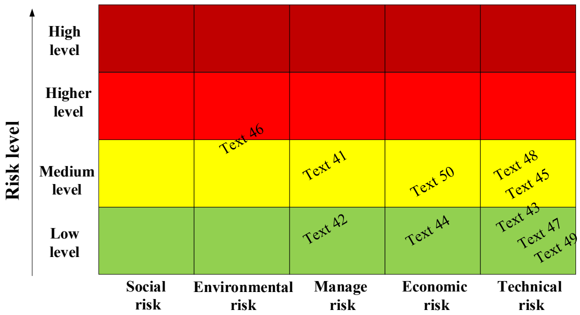

Figure 7. Risk decision matrix example diagram

5. Conclusion

This paper proposes a method of grid investment risk identification. The algorithm combines NLP technology and machine learning classifier, and through the learning of neural network, a classification model based on PNN and a risk degree evaluation model based on neural network are constructed to solve the problem of investment risk identification. The research results show that the risk identification error of the method constructed in this paper is within 5%, and the risk decision matrix can more intuitively show the degree of risk, and can more effectively propose risk response measures, thereby reducing investment risks to a greater extent.

ACKNOWLEDGMENT

Supported by the Science and Technology Project of State Grid Corporation of China(1400-202157215A-0-0-00).

REFERENCES

[1] MA Qian, WANG Zhaocong, PAN Xueping, LIU Xiaofan. Evaluation method of power grid investment decision based on utility function under new electricity reform environment [J].Electric Power Automation Equipment,2019,39(12):198-204.

[2] MA Yanfeng, YANG Xiaokuan, WANG Zijian, DONG Ling, ZHAO Shuqiang, CAI Yongqing. Operation risk assessment for power system with large-scale wind power integration based on value at risk.[J/OL]. Power System Technology:1-8[2021-03 -11].

[3] LI Yufeng, LIU Yuanchun, HU Dashan, GUO Jinyang, WANG Xueqiang. Study on risk classification assessment method of hydropower station project based on risk matrix-Taking Wudongde hydropower station as example[J]. Journal of Safety Science and Technology,2020,16 (01):130-134.

[4] WANG Zhaocong, MA Qian. Investment risk analysis of power grid enterprises based on hierarchy analysis and monte carlo method [J]. Smart Power,2018,46(07):42-48+74.

[5] ZHANG Jiaan，AN Shixing，CHEN Jian，et al．Distribution network risk assessment considering the flexible access of DG [J]. Distribution & Utilization，2019，36（5）：29-33.

[6] ZHOU Ming, DUAN Nan, LIU Shujie, SHEN Xiangyang. Progress in neural NLP: modeling, learning, and reasoning[J].Engineering,2020,6(03):155-188.

[7] ZHANG Lei, JIANG Yu, SUN Li. An improved TF-IDF text clustering method[J]. Journal of Jilin University (Science Edition),2021,59(05):1199-1204.

[8] WANG Jinheng, SHAN Zhilong, TAN Hansong, WANG Yulin. An improved TF-IDF text clustering method [J].Computer Science,2021,48(06):338-342.

[9] Salsabeel Shapsough, Rached Dhaouadi, Imran Zualkernan. Using linear regression and back propagation neural networks to predict performance of soiled PV modules[J]. Procedia Computer Science,2019,155.