WEP
Warwick
Evans
Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 11
EMFT 2024

Research on Predicting Enterprise New Quality Productivity Based on K-Means Clustering and BP Neural Network

Yifan Ding [1, #, *], Xingyu Zhou [2, #]

[1] School of Crowe Chinese Auditing, Nanjing Audit University, Nanjjng, China, 211815

[2] School of Government Audit, Nanjing Audit University, Nanjjng, China, 211815

* Corresponding Author Email: 18052911860@163.com

[#]These authors contributed equally.

Abstract. Firstly, using text mining methods and analyzing the frequency of words related to new quality productivity in enterprise annual reports, we calculated the new quality productivity index for each enterprise from 2007 to 2022. We constructed 25 predictive features from three aspects: enterprise digital transformation, ESG, and productivity. To improve the efficiency of machine learning methods, the Boruta algorithm was used for feature selection, resulting in 21 predictive features. To enhance the accuracy of enterprise clustering, principal component analysis (PCA) and UMAP methods were used for dimensionality reduction, with UMAP showing better results than PCA. Using the reduced-dimensional features, the K-means clustering algorithm was applied to cluster each enterprise. The BP neural network method was then used to predict the new quality productivity index, achieving an accuracy of 97.7%. The study concluded: 1) "Technology" appeared most frequently (544,610 times) in annual reports, underpinning new quality productivity development; 2) Digital transformation, ESG, and productivity significantly impact new quality productivity; 3) Enterprises with longer establishment times have higher productivity indexes, promoting high-quality development; 4) Financial characteristics and ESG indicators can predict a company's level of new quality development.

Keywords: New Quality Productivity, Boruta Algorithm, UMAP Algorithm, K-Means, BP Neural Network.

1. Introduction

Song Jia et al. [1] mentioned that the core of new quality productivity lies in innovation, with the main carrier being industry. Its nature is manifested in the emergence of a large number of disruptive innovative technologies, which release tremendous momentum through industrialization. In terms of quality, it is manifested in the rapid development of emerging industries with higher technological levels, better economic benefits, and more friendliness, thereby promoting high-quality economic development. Liu Caixia [2] believes that new quality productivity can be combined with green logistics, based on the new development pattern, and jointly promote the empowerment of green logistics by new quality productivity through sharing and building a green logistics ecosystem. Ma Rong [3] proposed that the construction of new digital infrastructure from the perspective of new quality productivity is not only a tool for promoting high-quality development, but also an important lever for achieving deep industrial transformation and upgrading. Zhang Jones and Chang Peiqi [4] believe that financial support for new quality productivity is strong. Huang Yong [5] proposed to carry out the "Artificial Intelligence+" action, create a digital industry cluster with international competitiveness, and become an important engine for accelerating the cultivation and development of new quality productivity. This article uses text mining methods to mine keywords in corporate annual reports and measure the company's new quality productivity. After using the Boruta model to reduce the dimensionality of the indicators. Innovatively using BP neural network models to predict the new quality productivity of enterprises from three aspects: the degree of digital transformation, ESG, and productivity. New quality productivity forecasting can help companies understand their future competitiveness in the industry in advance and make corresponding strategic plans and

adjustments. By better grasping their own and competitors' productivity levels, companies can more accurately predict their future market position, profitability, and prospects. In summary, predicting the new quality productivity of enterprises is of great significance for both the development of enterprises and the overall economy. Through accurate prediction, enterprises can better cope with market competition, improve production efficiency, enhance product quality, and achieve sustainable development. Meanwhile, policy makers can also formulate more targeted policies based on the predicted results of new quality productivity, guiding industrial upgrading and economic development.

2. ESG indicator construction

This article refers to the Wind ESG evaluation system to score corporate ESG indicators. The Wind ESG rating provides a predictive assessment of a company's substantial ESG risks and its ability to sustainably operate, measuring the company's commitment and performance in ESG, and helping investors identify important risks and opportunities in their investments [6]. At present, it has covered all A-share and Hong Kong listed companies and important bond issuers, with about 8000+companies.

The Wind ESG evaluation system consists of management practice evaluation and dispute event evaluation, which can comprehensively reflect the ESG management practice level and major unexpected risks of enterprises, and combine with the development of China's capital market, regulatory policies, and company ESG practices to form a localized characteristic indicator system, which can be more scientifically applicable to Chinese companies. The scope of ESG assessment covers 3 dimensions, 27 topics, and 300+data points. The Wind ESG indicator system is shown in Table 1:

Table 1. Wind ESG Indicator System.

Wind ESG Indicator System		
/	Hire	/
Environmental Management	Occupational Health and Safety Production	Corporate Covernance
Energy and Climate Change	Development and Training	ESG Governance
Water Resource	Research and Innovation	Dong Jiangao
Raw Materials and Waste	Supply Chain	Equity and Shareholders

2.1. Construction of productivity two factor indicators

Based on the theory of the two factors of productivity, construct an indicator system. Productivity includes two elements: labor force and production tools. Among them, labor force consists of two sub elements: active labor and materialized labor (labor object); Production tools consist of two sub factors: hard technology and soft technology. Considering the innovative connotation of new quality productivity, the indicators of active labor sub factors are measured by R&D personnel salary, R&D personnel proportion, and high education personnel proportion, respectively; The indicators of physical labor sub factors are represented by the proportion of fixed assets. Considering that enterprises with new quality productivity are mainly concentrated in the high-precision technology field of equipment manufacturing, most of these enterprises rely on high-end machinery and instruments for production, and machine production replaces human labor[7]. The proportion of manufacturing costs for these enterprises is higher than that of other enterprises, so the proportion of manufacturing costs is also included in the indicator selection. The main sub factors of hard technology are hardware equipment related to R&D investment, so they are measured by the proportion of direct R&D investment, depreciation and amortization, and leasing expenses, respectively. At the same time, considering the role of intangible assets such as software, they are also measured by the proportion of intangible assets; The sub factors of soft technology mainly include total asset turnover and equity multiplier to measure. Considering that the higher the equity

multiplier, the higher the financial risk of the enterprise, this indicator is a negative indicator that is inconsistent with other indicators. Therefore, the reciprocal of the equity multiplier is used to represent it. The higher the reciprocal, the lower the risk, indicating that the productivity level of the enterprise is better.

2.2. Descriptive analysis of data

In order to make better statistical inferences about the data, the numerical characteristics of the relevant indicators of enterprise digital transformation were first analyzed, as shown in Table 2 below.

Table 2. Descriptive statistical results of major variables.

Variable Name	Observations	Mean Value	Standard Deviation	Minimum Value	Median	Maximum Value
New Quality Productivity	32806	6.9654	6.581	0.97	5.45	66.29
Degree of digital transformation	32806	1.6737	1.460	0.00	1.61	5.63
Big data technology	32806	1.4678	1.266	0.00	1.39	5.07
Big data technology	32806	1.7854	1.235	0.00	1.79	5.24
Big data technology	32806	1.6025	0.961	0.00	1.61	4.26
Cloud computing technology	32806	2.2580	1.084	0.00	2.26	4.47
Blockchain technology	32806	3.1516	0.525	1.01	3.23	4.17
Intelligent manufacturing	32806	3.1881	0.408	1.30	3.21	4.07

Table 2 shows the descriptive statistics of the main variables. The mean, median, and standard deviation of the key variable 'new quality productivity' are 6.9654, 5.45, and 6.581, indicating a normal distribution of innovation levels and meeting the experimental conditions. At the same time, there is a significant difference between the minimum and maximum values, with the mean being much smaller than the maximum value, indicating significant differences in innovation levels among different enterprises. Each control variable can play a good controlling role in this study.

Table 3. Descriptive Statistical Results of ESG Indicator System.

Variable Name	Observations	mean value	standard deviation	minimum value	median	Maximum value
E_Score	32806	15.5002	17.345	0.00	8.55	86.71
S_Score	32806	25.3704	12.007	1.74	24.40	63.74
G_Score	32806	25.2055	10.142	2.62	23.87	57.83

Table 3 shows that the explanatory variable ESG follows a normal distribution, and there is a significant difference between the minimum and maximum values, indicating that the sample coverage selected in this study is relatively wide, which also promotes the representation of research conclusions.

Table 4. Descriptive statistical results of the new productivity indicator system.

Variable Name	Observations	mean value	standard deviation	minimum value	median	Maximum value
Other	34354	0.0144	0.023	0.00	0.01	0.19
Inv	34354	0.1402	0.129	0.00	0.11	0.78
ROA	34354	0.0363	0.066	-0.47	0.04	0.22
Loss	34354	0.1158	0.320	0.00	0.00	1.00
Lev	34354	0.4149	0.207	0.03	0.40	0.93
Size	34354	22.1830	1.301	19.57	21.98	26.45
Age	34354	10.7267	7.803	1.00	9.00	30.00

Table 4 shows that the data of various financial indicators basically follow a normal distribution, and there are significant differences in the minimum and maximum values of a series of control variables.

3. Feature selection based on machine learning methods

3.1. Feature selection based on Boruta algorithm

In order to provide a predictive model for new production capacity, machine learning methods are first used for feature screening [8]. The Boruta algorithm is based on random forests and effectively selects important features from a large number of features by comparing the importance of real features and randomly arranged features, thereby improving the prediction accuracy of the model, reducing the risk of overfitting, and lowering computational costs. The feature selection results of Boruta algorithm are shown in Figure 1-2. Each line in Figure 1 represents a predicted feature, with the vertical axis indicating the importance of the feature. The blue line serves as the filtering line, the red line represents the "rejected" and "hesitant" features, and the green line represents the "acceptable" features.

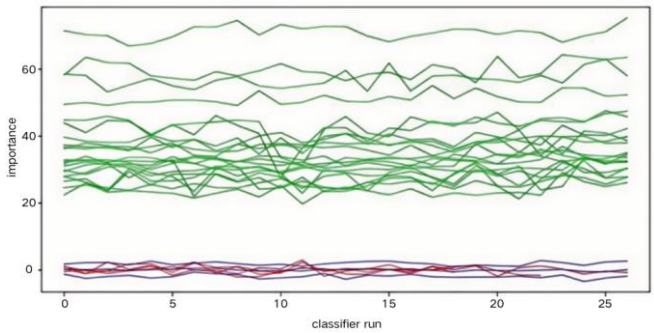

Figure 1. Feature screening structure of Boruta algorithm.

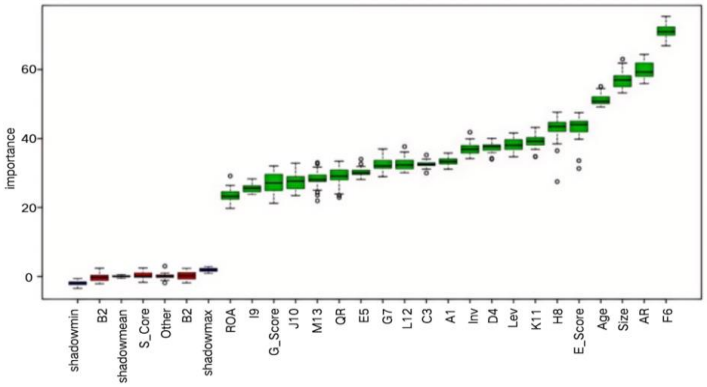

Figure 2. Feature screening results of Boruta algorithm.

From Figures 1 and 2, it can be seen that after 100 iterations of Boruta algorithm training, 21 out of the 25 feature variables mentioned above were accepted, 1 feature was rejected, and 3 features were classified as "hesitant". Among them, S-Score, Other, and Loss are the "rejected" feature variables; B2 is the characteristic variable of hesitation. From this, it can be concluded that big data technology for enterprise digital transformation, social scores in ESG scores, and other comprehensive benefits and losses in productivity factors have little effect on promoting new quality productivity. This article will remove these three indicators and use the remaining 21 indicators as features for machine learning.

3.2. K-means clustering

By conducting cluster analysis on indicators, indicators with similar characteristics can be grouped together to better identify patterns and patterns in the data. This helps identify potential subgroups in the data, discover correlations between variables, and provide guidance for subsequent feature engineering and model selection [9].

This article uses unsupervised learning K-means algorithm to cluster and analyze enterprises based on 21 relevant indicators of new quality productivity, in order to understand the internal structure of data, discover potential outliers and important features, reduce data dimensionality and noise, improve the robustness and predictive ability of the model, and provide useful guidance for subsequent feature engineering and model training. Firstly, using the elbow rule to determine the number of cluster centers, the results are shown in Figure 3.

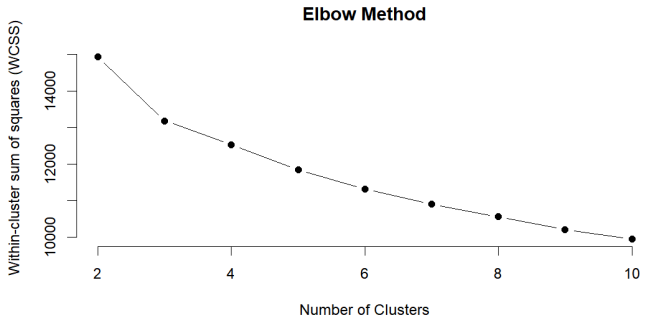

Figure 3. Analysis of 21 Indicator Elbow Rules.

From Figure 3, it can be seen that there are 9 clusters based on the elbow method. Due to the consideration of 21 features at this point, it may have multiple impacts on the effectiveness of K-means clustering: too many indicators will increase the possibility of noise and redundancy in the data. Some indicators may be highly correlated, resulting in redundant information, while some indicators may be noise [10]. Adding these indicators may interfere with the learning process of the clustering model. Excessive indicators will increase the dimensionality of the feature space, making it more difficult for clustering algorithms to find a suitable partition to ensure high-quality clustering results. Excessive indicators consume too much space, which can lead to an increase in the spatial complexity of clustering results and a relative blurring of clustering divisions. The more indicators there are, the higher the complexity of calculating distance and similarity, which greatly increases the computational cost of K-means. Therefore, in order to enhance the clustering performance of the K-means clustering algorithm. This article performs dimensionality reduction on the indicators before conducting cluster analysis.

4. Cluster analysis of new quality productivity based on dimensionality reduction method

4.1. Cluster analysis based on principal component analysis method

Firstly, conduct PCA analysis on the 21 indicators initially selected by the Boruta algorithm. The cumulative contribution obtained is shown in Figure 4:

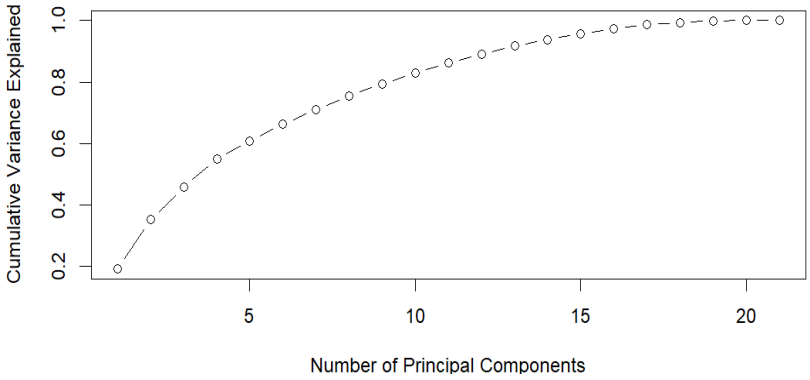

Figure 4. PCA Cumulative Contribution Chart.

According to Figure 4, the cumulative contribution of the first fourteen features has reached 90%. Then, based on these 14 features, K-means clustering analysis was performed, and the results are shown in Figure 5-6.

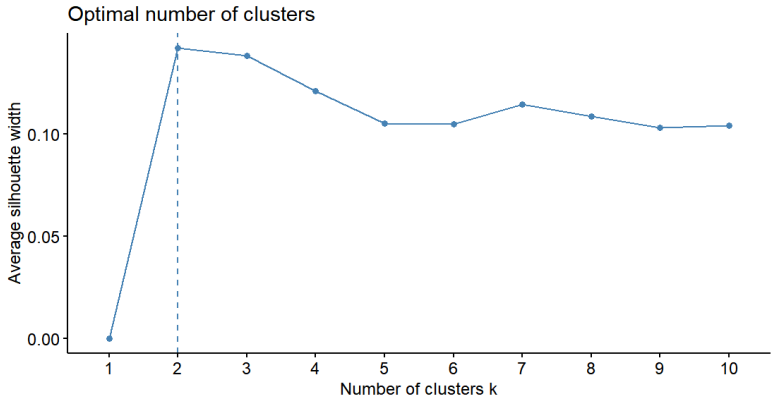

Figure 5. Contour coefficients of the first 14 indicators.

The highest point of the curve appears when the number of cluster centers is 2, and the highest value of the contour coefficient is around 0.15, which results in poor performance.

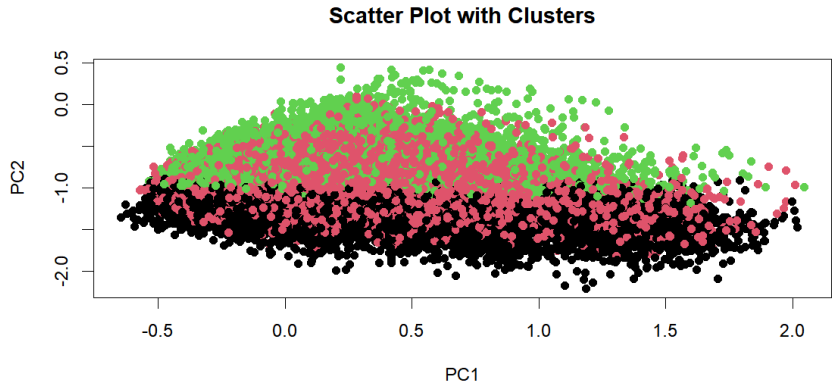

Figure 6. Cluster diagram of the first 14 indicators.

The clustering effect diagram is shown in Figure 6, where it can be seen that there is a clear intersection between each class. According to the two figures above, it can be seen. After using PCA dimensionality reduction processing, the clustering effect of K-means algorithm is still not good.

4.2. Cluster analysis based on UMAP algorithm

The advantage of UMAP in feature selection lies in its non-linear nature. Compared to PCA's linear projection, UMAP can better preserve the nonlinear structure of data and capture the complex relationships between features when integrating and filtering features. This enables UMAP to perform better in clustering and distribution of features. At the same time, by converting the K-means

clustering results into binary classification tasks, the clusters classified into the first category were taken as positive examples, and the other clusters were taken as negative examples. The similarity between clusters in the clustering results was used as the classifier's prediction probability, and the ROC curve was calculated. The ROC curve shows the relationship between true case rate and false positive case rate in the graph, where true case rate represents the ability to correctly identify positive cases, and false positive case rate represents the ability to incorrectly identify negative cases as positive cases. The drawing of ROC curve can help us understand the classification ability of K-means clustering results and the performance of classifiers at different thresholds, providing important references for further model optimization and application.

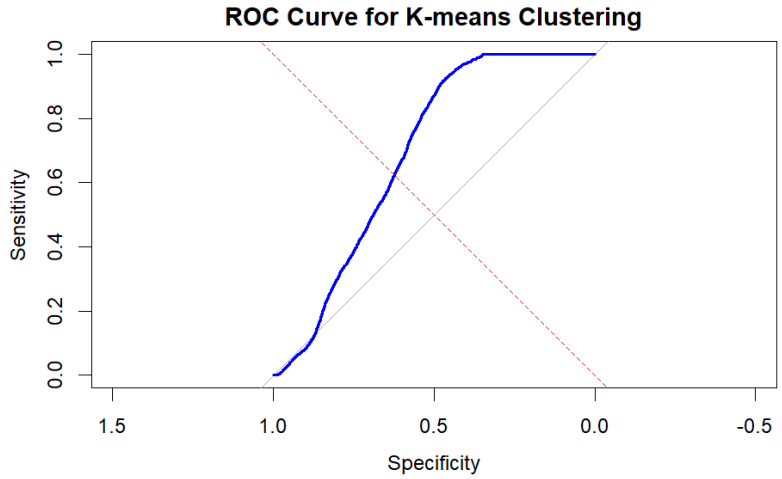

Figure 7. ROC rendering.

The area under the ROC curve (AUC) is a commonly used performance metric to evaluate the quality of classifiers, with values closer to 1 indicating better classifier performance. As shown in Figure 7, the AUC value of the K-Means clustering model after UMAP dimensionality reduction is greater than 0.5 and very close to 1. From the ROC curve, it can be concluded that the K-means clustering model has good stability on different datasets after UMAP dimensionality reduction.

The results of dimensionality reduction using UMAP are shown in Figure 8, where all indicators are clearly classified into three main categories.

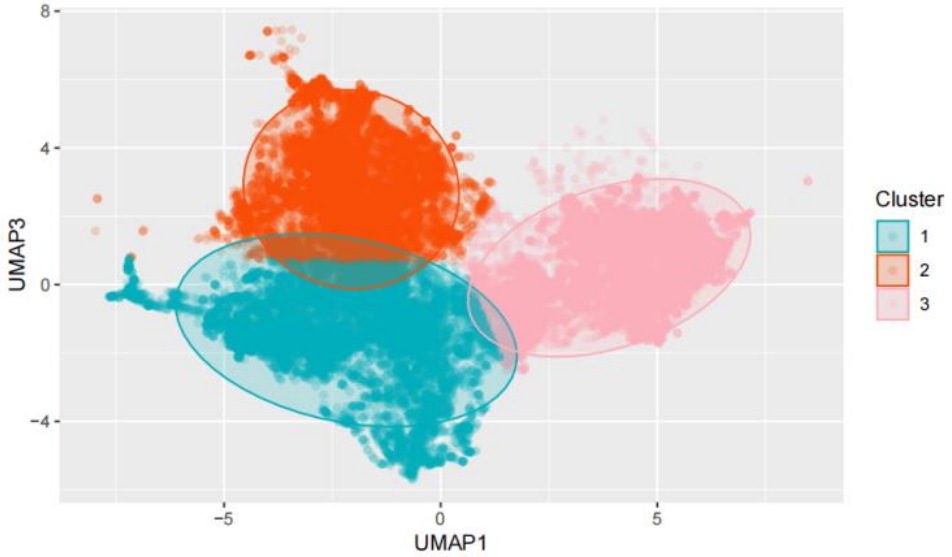

Figure 8. UMAP dimensionality reduction effect diagram.

UMAP reduced the values of the original 21 indicators to 3 indicators, named UMAP1, 2, and 3 respectively. At the same time, the UMAP algorithm reduces the dimensions of 21 indicators to specific values M for 3 indicators.

After dimensionality reduction using UMAP, cluster again using K-means algorithm.

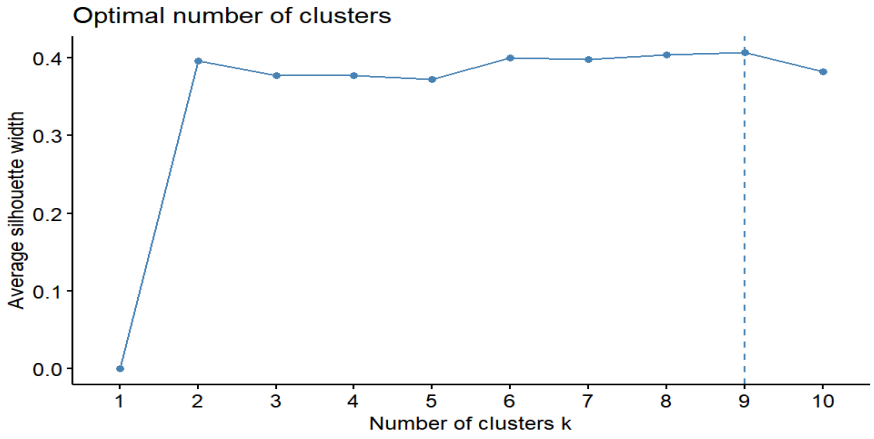

Figure 9. UMAP reduced dimensional contour coefficient graph.

As shown in Figure 9. Compared to the clustering effect after PCA dimensionality reduction, the highest value of contour coefficient after UMAP dimensionality reduction reached 0.4, indicating excellent clustering performance.

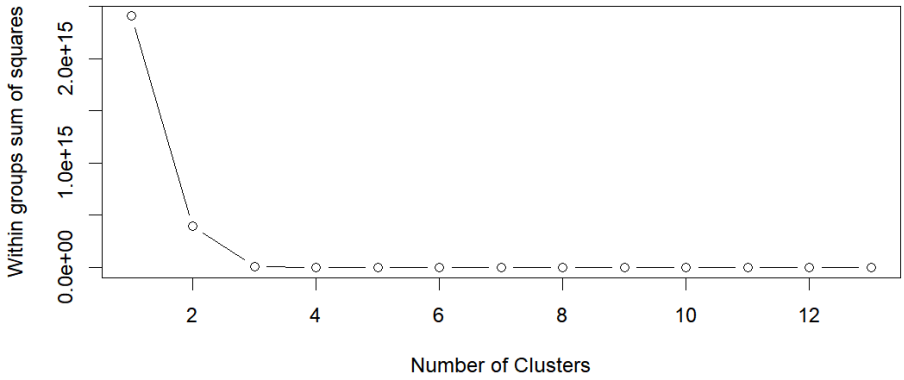

Figure 10. Elbow Rule Diagram after UMAP Dimensionality Reduction.

From Figure 10, it can be seen that according to the elbow rule, when the number of cluster centers is 3, the amplitude of straight turns is the largest. So confirm that the number of cluster centers is 3. At the same time, when the number of cluster centers is 3, the value of the contour coefficient is also relatively large. So the final confirmed number of cluster centers is 3.

5. Conclusions

According to the text mining model, it can be inferred that 'technology' has the highest frequency of words in corporate annual reports. This indicates that continuous technological progress and innovation can drive enterprises to adopt new technologies, processes, and equipment, thereby improving production efficiency and quality, reducing costs, and achieving growth in new quality productivity. Digital transformation can improve the production efficiency and flexibility of enterprises, promote the optimization and innovation of business processes, and thus drive the improvement of new quality productivity. The sustainable development of the environment, society, and corporate governance is closely related to the new quality productivity of enterprises. Productivity improvement is the fundamental element for enterprises to achieve new quality productivity growth. Enterprises that have been established for a longer period of time may indeed have certain advantages in terms of new quality productivity. Over the long term, these enterprises may have accumulated rich experience and resources, established relatively complete operational systems and management mechanisms, and thus be more capable of responding to market changes and challenges, promoting the improvement of new quality productivity. Based on the financial characteristics and ESG (Environmental, Social, and Governance) indicators of a company, it is

possible to predict its level of new quality development. Financial characteristics mainly include indicators such as a company's profitability, debt-paying ability, and growth potential, while ESG indicators cover a company's performance in environmental, social, and governance aspects. Combining the two can provide a more comprehensive assessment of a company's development potential and long-term value.

References

[1] Gao Pengli, Ren Dalu, Li Chaohui, et al. Spatial distribution prediction of soil organic matter based on Boruta algorithm and GA optimized mixed geostatistical model [J]. Geophysical and Chemical Exploration, 2024, 48 (03): 747-758

[2] Yu Shiao, Kong Wei, Ma Rujia, etc Photon Counting Lidar Point Cloud Filtering Based on BP Neural Network [J/OL] Progress in Laser and Optoelectronics, 2024,: 1-15

[3] Song Caizhu, Tana, Yan Caixia, etc Prediction Model and Application of Environmental Factors in Sunlight Greenhouse: Based on BP Neural Network [J] Agricultural Mechanization Research, 2024, 46 (10)

[4] Wang Dong, Yang Yuxin Train wheel tread wear prediction model based on PCA-MSSA-BP neural network [J] Technological Innovation and Application, 2024, 14 (12): 49-54

[5] Huichangwu, Xu Dejie, Gong Liang, etc Research on the Evaluation Index System of Urban Rail Transit Operation Safety Based on Principal Component Analysis [J] Urban Rapid Transit, 2024, 37 (02): 131-138

[6] Liu Junli, Miao Bingrong, Zhang Ying, etc A fault feature extraction method for rolling bearings based on improved VMD and UMAP [J] Mechanical Transmission, 2023, 47 (06): 130-138

[7] Chen Jiayuan, Zhang Lu Research on the Evaluation of Urban Logistics Competitiveness and Optimization Path of Logistics Network Based on Principal Component Analysis: A Case Study of Guangdong Province [J] Journal of Shanghai Institute of Economic Management Cadres, 2024, 22 (02): 28-45

[8] Zhao Chenyu, Wang Wenchun, Li Xuesong. How Digital Transformation Affects Total Factor Productivity of Enterprises [J]. Finance and Trade Economics, 2021, 42 (07): 114-129

[9] Wang Qisheng, Xiong Junnan, Cheng Weiming, etc A landslide susceptibility evaluation method combining statistical methods, machine learning models, and clustering algorithms [J] Journal of Earth Information Science, 2024, 26 (03): 620-637

[10] Jinyang Exploration of the Applicability of Principal Component Analysis in the Monitoring and Analysis of Hydroelectric Unit Operation [J] Energy Engineering, 2024, 44 (01): 79-84