Application of Big Data to common statistical methods based on game systems

ABSTRACT

In this paper, statistical methods are adopted to analyze its role in data processing and prediction, emphasizingthree statistical models of time series model, neural network model based on genetic algorithm and K-means ++ algorithm, systematically introducing how to classify and predict data based on three common statistical methods, and find out the characteristics of data by using the trend of data change. In order to grasp the actual user and other relevant information. Taking MCM C project in 2023 as an example, this paper classifies and analyze the given data, find out the variation characteristics of variables, predicts the development and changes of variables, and provides direction for game companies to further strategies, so as to enhance users' sense of experience. Finally, through comprehensive analysis and processing, three kinds of prediction and classification statistical models are applied deeply to cope with this practical issue.

Keywords:Time series analysis, BP neural network, K-means clustering analysis, correlation analysis

1. INTRODUCTION

With the development of society, economy and science and technology, statistics plays an increasingly significant role in modern management and social life. In addition, statistics have been widely used in medical research, market research, financial analysis, educational evaluation and other fields, and has important significance and status. For example, conducting quantitative research on residents' consumption patterns, predicting future market trends and risks, and so on.

There are many statistical models, such as logistic regression, the decision tree,the cluster analysis, principal component analysis, time series model, multiple regression analysis, neural network, random forest and so on.The standpoint of statistical model is applied to all aspects of social life, but the traditional single linear modelto tackle the complex issue is not well.This paper will take 2023 MCM C Project-Wordle as an example, emphasizing on the application of statistical models in the analysis of game data, focusing on the study of time series model, neural network model based on genetic algorithm and the analysis and processing of data by K-Means ++ algorithm.

Among them, the time series analysis has some advantages in terms of processing speed, can effectively predict the development and change of variables. It has a high level of data fitting and accuracy.The weight and threshold of the hidden layer in BP neural network is optimized, and the neural network model based on genetic algorithm is established to improve the learning rate and the prediction accuracy is high.Applying the K-Means ++ algorithm to the clustering model can maximize the similarity between the same types and the difference between the different types, and the clustering effect is good. Furthermore, the application of these statistical methods in the processing of game data is improved, and the development and optimization of the following help is provided.

2. THE TIME SERIES ANALYSIS METHOD

Time series analysis method prefers to arrange a group of observed valuesof the same variables such as economic development, purchasing power and sales change in a chronological order to form a statistical time series, and subsequently use certain digital methods to make it extend outward to predict the development trendof the market and determine the predicted value of the market. It is a method to forecast data[1].

2.1Data Preprocessing

Based on the data given in Question C of MCM, do data pre-processing first, and invalid data and unrepresentative data are deleted to obtain 342 sets of valid data. The specific process is shown in the figure below:

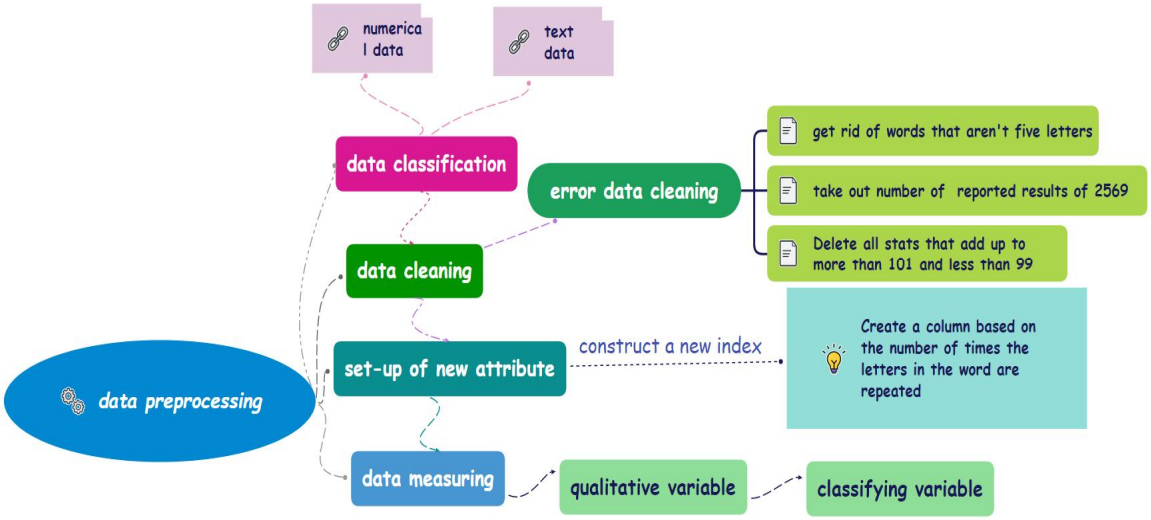


Figure 1 data pre-processing

2.2 Establish ATime Series Prediction Model

It is found that general regression analysis can not effectively predict the factors for which no data can be obtained[2]. Therefore, as for the prediction issue of the number and change of results reported by game users in question C, we choose the time series model to solve such issues well. Construct a time series graph, study the trend of the image, and analyze how the number of reported results changes over time. Combined with the data characteristics, the most appropriate time series prediction model is selected and time is taken as the independent variable to judge its trend, and subsequently the consequence value of March 1, 2023 is predicted.

Define the time and date, the first case "January 7, 2022" is defined as the first day, the second case "January 8, 2022" is defined as the second day, and so on, all the way to the 342 sets of data, thus accomplishing the definition of the time variable. Therefore, "March 1, 2023" which needs to be predicted is the 402nd day. By drawing the time series diagram of the original data (Figure 2), it is observed that the data contains a linear trend and does not show the data characteristics transforming with the seasons, so there is no seasonal fluctuation. Now, "Number of reported results" is set as the dependent variable, and the non-seasonal exponential smoothing method is applied to the analysis of time series data based on SPSS, which canbecome better.

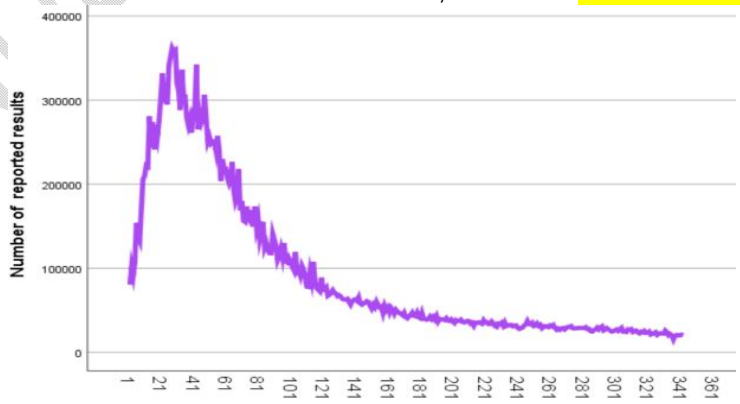


Figure 2. Raw data graph

The non-seasonal time series prediction models mainly include simple models, Holt linear trend model, Brown linear trend model and damped trend **mode**[3]. These four methods are used separately to forecast the number of reported results on March 1, 2023. The fitting results show that,

as shown in Table 1, the stationary R squared under the Holt linear trend model is the largest. It can be seen that the prediction results of the Holt linear trend model are more accurate[4]. Therefore, this method is selected for data prediction and the model analysis.

Table 1. R squared of the four methods

Types of models	Stable R square	R square
Naive model	.127	.982
Holt Linear trend model	.713	.984
Brown Linear trend model	.696	.981
The damping trend model	.231	.984

.The paragraph suggests that the time series prediction model is not only effective in accurately fitting the original quantity data from the report but also capable of making relatively accurate predictions for future quantities. The prediction results for days 343 to 402 are presented in Table 2.

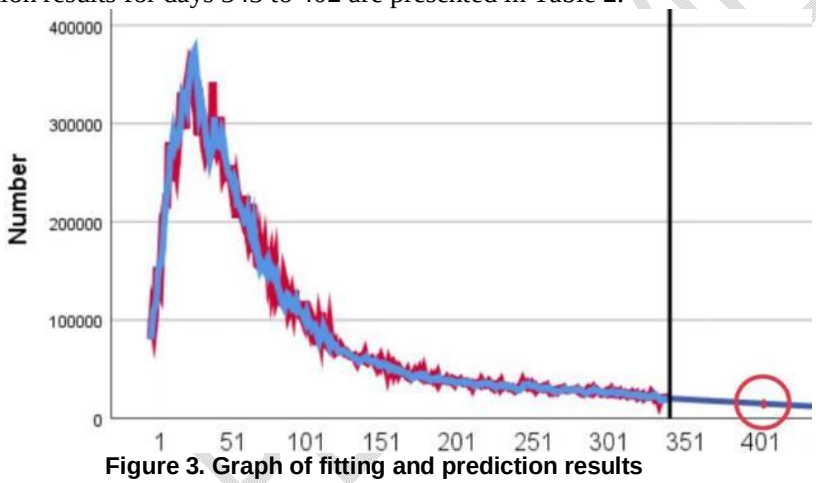


Figure 3. Graph of fitting and prediction results

In order to further verify the stability and rationality of the model, the stationary R-square and R-square values and the significance level of the goodness-of-fit measurement indexes in the model statistics were selected to test the model. The statistical table of model fitting degree is shown in Table 2. From Table 2, stationary R squared is 0.713, R squared is 0.984, and the value is close to 1, indicating that the model has a good fitting effect. The significance level was obviously less than 0.05, indicating that the significance was good. The above indicates that the time series prediction model has high fitting degree and high accuracy of prediction results.

Table 2. Statistical table of model fit degree

Model	Model fit degree statistics		Yang Box Q(18)		
Number of predictive variables	Stationary R square	R-squared	Statistics	DF	Significance Level
0	0.713	0.984	77.386	16	0.000

3. BP NEURAL NETWORK OPTIMIZED BY GENETIC ALGORITHM METHOD

3.1 BP Neural Network Prediction Model

BP neural network is a multilayer feedforward network trained by error BP algorithm[5]. The network constantly adjusts the network weight and threshold through back propagation to minimize the total square error of the network. The topology structure of BP neural network model includes the input layer, hidden layer and output layer[6]. Each layer has many "neurons", and the output from the

previous layer is used as input to the next. A neuron needs four parameters and a mapping function to get its output. BP neural networks have a training set for training the network and a verification set for validating the network. Therefore, BP neural network algorithm has good adaptability and classification recognition ability.

3.2 The Treatment Of Words

If you can count the stars, you can know the chess position. The prediction of the proportion of different word attempts can be regarded as a complex function mapping problem with five letters as input and the proportion of attempts as output[7]. Fortunately, we know that letters and numbers can establish a one-to-one correspondence with ASCII codes[8]. We can extract the five letters of the word and convert them into five numbers, so that each word will be a vector with five elements and treat all letters as lowercase letters, and let "a" be "1", and so on (Table. 3).

Table. 3. Word digitization Result Graph (Part)

manly	13	1	14	12	25
molar	13	15	12	1	18
havoc	8	1	22	15	3
impel	9	13	16	5	12
condo	3	15	14	4	15
judge	10	21	4	7	5
extra	5	24	20	18	1
poise	16	15	9	19	5

3.3 The Prediction Results Of The BPNeural Network.

The number n_2 of the hidden layer neural network generally has the following approximate relationship with the number n_1 of input layer neurons:

$$n_2 = 2n_1 + 1 \tag{1}$$

That is to say, the number of neurons in our hidden layer is 11.

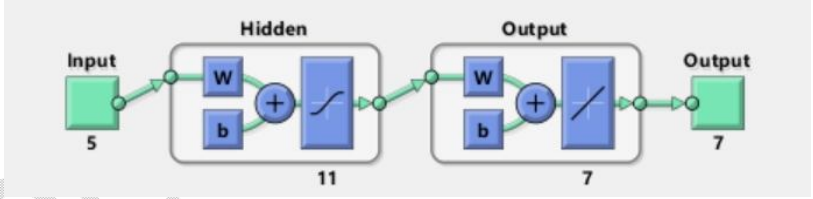


Figure 4. Schematic diagram of neural network structure

The transfer function of the neurons in the hidden layer of the neural network uses the S-type logarithmic function "logsig", and the transfer function of the neurons in the output layer uses the linear function "purelin".

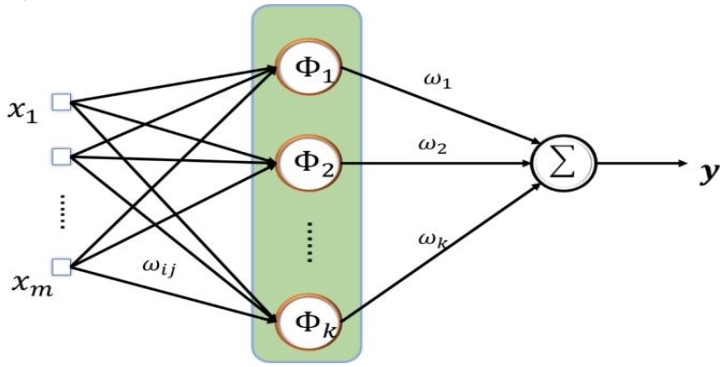


Figure 5. BP neural network algorithm flow diagram

Therefore, we can get the relationship between the input layer and the output layer of the network:

$$\hat{y} = \sum_{j=1}^{r} v_j f\left[\sum_{i=1}^{m} w_{ij} P_i + \theta_j\right] (k = 1, 2, \ldots\ldots N) \tag{2}$$

MATLAB neural network toolbox was used for fitting, and the fitting performance was as follows:

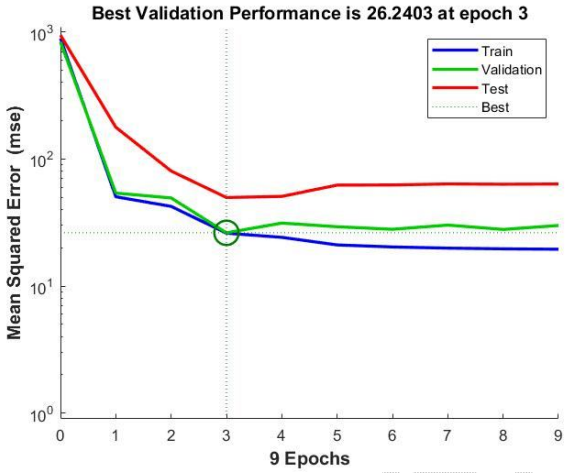


Figure 6. Training performance of neural networks

In fact, although the mean square error decreased rapidly in the early stage, we can see that with the deepening of training, the gap between the trained network mean square error became larger and larger. At the same time, the mean square error of the test set did not continue to decline or stabilize. Whether such a model is acceptable remains to be seen. If possible, we should optimize and adjust the traditional BP neural network prediction model as much as possible in order to improve the predictedeffect of the model.

3.4 Optimization Model Of BP Neural Network Based On Genetic Algorithm

The initial weights and thresholds of BP neural network training are generally generated randomly. In the process of BP neural network training, the algorithm is used to update the weight threshold through forward propagation data and reverse error transmission, and then the gradient descent method is used to update the parameters[9]. In fact, we can use some algorithms to determine the appropriate values of these parameters, such as genetic algorithms.

There are several reasons why one might choose to use the Genetic Algorithm (GA) to optimize BP neural networks. The GA is a population-based optimization technique that can search the solution space more effectively compared to traditional optimization methods[10]. It has the ability to explore a wide range of solutions and converge towards the global optimum, rather than getting stuck in local optima. The GA can serve as a regularization technique to prevent overfitting by maintaining diversity in the population through genetic operators like crossover and mutation. The GA is suitable for optimizing BP neural networks when the objective function is non-differentiable or contains discontinuities, as it doesn't rely on gradients like traditional methods. Moreover, using genetic algorithms instead of gradient descent can greatly improve efficiency.

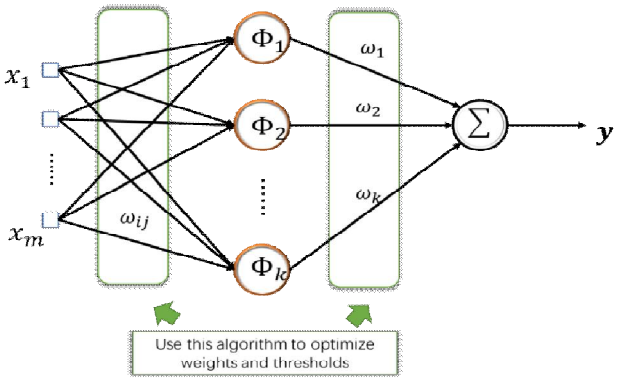


Figure 7. Schematic diagram of BP neural network algorithm optimized by genetic algorithm

The optimization of BP neural network by GA means that the initial weights and thresholds of BP network are optimized by GA, and the optimized BP network can obtain better predictive output[11]. The main steps of genetic algorithm GA to optimize BP network are as follows:

Figure 8. The steps of genetic algorithm GA to optimize BP network

Step1 Population initialization: All individuals are encoded by real numbers, and each individual contains all network weights and thresholds. This ensures that we can completely identify a network through input and output. Virtually all of our information is already digital.

Step2 Construct the fitness function: Genetic algorithm in the final analysis is an optimization algorithm used to find the extreme value, so naturally there needs to be a function to be optimized to represent some indicators of the neural network. Firstly, the initial weight and threshold of the BP neural network algorithm can be obtained by each individual, and then the general BP neural network training is carried out. The prediction error norm of all prediction samples is taken as the fitness function, and the calculation formula is as follows:

$$F = k\sum_{i=1}^{n} | y_i - o_i | \tag{3}$$

Where n is the output of the network node, y_i is the expected output of the ith node of the BP neural network, o_i is the predicted output of the ith node, and k is the coefficient.

Step3 Selection operator: In the cross operation, the ergodic sampling method is adopted. In the cross operation formula (4), a_k is the kth chromosome and a_1 is the lth chromosome, where b is any number between 0 and 1,

$$\begin{aligned} a_{kj} &= a_{kj}(1-b) + a_{lj}b \\ a_{lj} &= a_{lj}(1-b) + a_{kj}b \end{aligned} \tag{4}$$

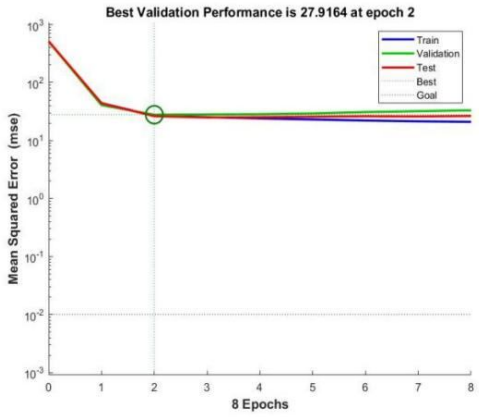
Step4 Selective variant gene: The random number generated by randomness is used to determine the variation and select the variant gene. If the selected gene code is 1, it becomes 0; otherwise, it becomes 1. Some basic parameters of genetic algorithm operation are shown in Table 4:

Table 4. Basic parameter table in genetic algorithm

population size	Maximum genetic algebra	Individual length	generation gap	Cross probability	Probability of variation
40	10	10	0.95	0.7	0.01

3.5 The Results Of BP Neural Network Prediction Model Optimized By Genetic Algorithm

With BP neural network optimized by genetic algorithm, we can get the fitting performance as shown in Figure 9. By comparing the degree of fitting between Figure 8 and Figure 11, we find that the prediction performance of BP neural network optimized by genetic algorithm has been significantly improved, and the mean square error of training set and test set is basically consistent. The genetic evolution process of genetic algorithm is shown in Figure 10.



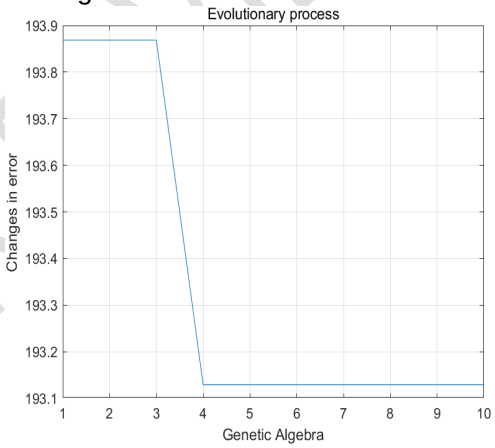


Figure 9. BP neural network training effect **Figure 10. Evolution ofgenetic algorithm**

The regression effect of the BP neural network model optimized by genetic algorithm is predicted. At 95% confidence, the correlation coefficient R is 0.92239, and the data always keep the trend of converging to the straight line.

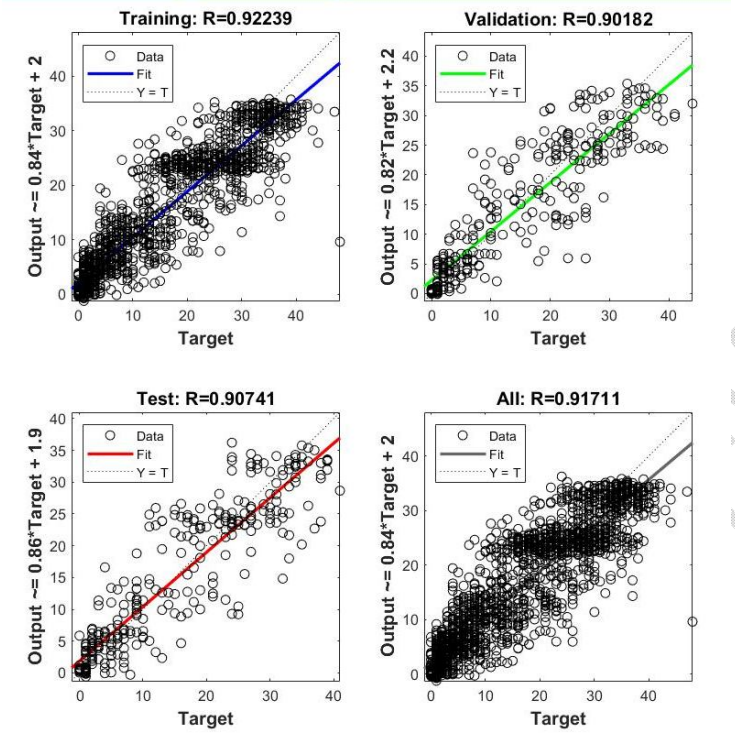


Figure 11. The fitting results of BP neural network optimized by genetic algorithms

Using the BP neural network optimization model based on genetic algorithm, the word " EERIE" is predicted. The result of word digitization is that the variable value of the word is defined as [5,5,19,15,5] and then substituted into the model to obtain the predicted value of percentages of (1, 2, 3, 4, 5, 6, X). (After rounding)

Table 5. Percentage predicted value

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
0%	5%	23%	34%	24%	12%	3%

4 K-MEANS ++ ALGORITHM METHOD

4.1 Construction Of K-means Cluster Analysis Model

K-means clustering analysis method is a kind of iterative solving clustering analysis algorithm[12]. The algorithm randomly selects K objects as the initial clustering center, then calculates the distance between each object and each seed clustering center, and assigns each object to the nearest clustering center .

According to the word attributes, the number of repeated letters in the word and the types of parts of speech included in the word, combined with the percentage of players' problem-solving times, the K-means clustering analysis method was used to analyze the data, and the classification results were summarized, and the similarities of the same class and the differences between different classes were found to form the basis and method for the classification of the difficulty of words[13].

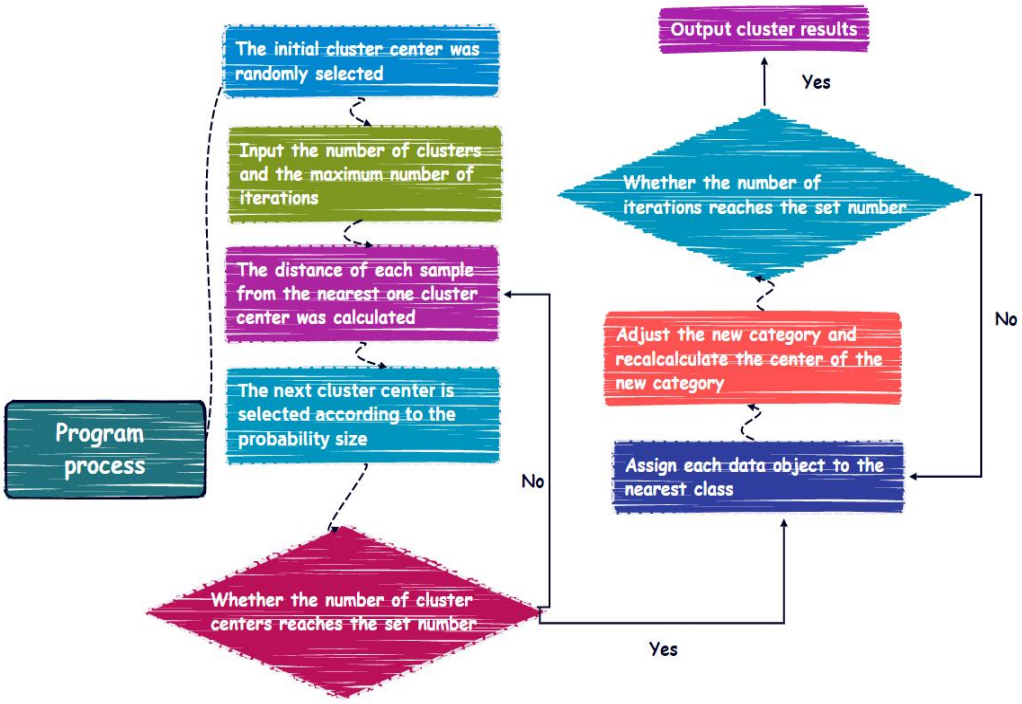


Figure 12. K-means cluster analysis flowchart

4.2 Classification Of Difficulty Of Words

The word attribute used in this question is the number of repeated letters in the word and the type of speak the word the word has. According to the number of players' problem solving and the word attribute, the difficulty of the word is divided by k-means clustering analysis method[14]. The K value is set as 2 by SPSS software, and the output result is shown in Table 6.

Therefore, according to the comparative analysis of the word attributes in classified words, the class basis of the difficulty of words can be obtained, and the name of each category is defined according to its characteristics. The first category is difficult, and the second category is simple. The specific differences in each category are as follows.

First type (difficulties): words with repeating letters, words with no repeating letters but containing two parts of speech, and words with no repeating letters but containing three parts of speech.

Second type (simple): words with no repeated letters and only one part of speech.

This kind of classification makes sense. When there are a lot of repeated letters in a word, it leads to a wider range of words that the player can think of, which decreases the chance that the player can complete it in a few times, and increases the chance of failure. At the same time, when there are multiple parts of speech, the player cannot determine the position of each word, resulting in too many attempts of the player, which exceeds the 6 chances set by the game rules, thus making it difficult for the player to guess the right word.

The results obtained from problem two are substituted into the model, and described, so as to complete prediction of the difficulty of "EERIE". Through the application of K-means clustering analysis method, the category of "EERIE" is judged by observing the changes in the number of various categories. According to the result (Table 6), the number of the first category is increased by one, so it is predicted that

"EERIE" is a word in the difficult category, and it is difficult to solve this word within 6 times. In terms of the word itself, the word's contains the repeated letter "E", and it is repeated forthird times, which accords with the classification standard set by us, and further indicates that the model constructed by this question has practical application value.

Table 6. Comparative analysis of the number of cases in each cluster

	classification	quantity
cluster	difficulty	141.00
	Simpleness	201.00
effect		342.00
deletion		.00

	classification	quantity
cluster	difficulty	142.00
	Simpleness	201.00
effect		343.00
deletion		.00

4.3 Cross Tabular Chi-squareTest

Cross-tabular Chi-square test is a correlation analysis method that can examine the correlation between disordered classification variables[15]. This method is used to test the degree of correlation between the classification of word difficulty, the number of repeated letters in the word and the class of parts of speech of the word respectively, so as to test the accuracy of the model.

Firstly, the two categories of difficulties and simplicity were quantified, and the difficulty was defined as "1" and the simplicity as "2". The cross-tabulation chi-square test was carried out by SPSS software, and the significance values were all less than 0.05. (Table 7) Thus, we can find that the two attributes of a word, namely the number of repeated letters in a word and the type of speak a word a word has, are significantly correlated with classification, and the number of repeated letters in a word is more significantly correlated with classification.

Table 7. Crosstab chi-square test 1

	Numerical value	df	significance
Pearson chi-square	114.432a	3	0
likelihood ratio	138.525	3	0
Linear association	95.076	1	0
Number of effect	343		

5.CONCLUSION

This paper focuses on the application of three statistical methods to game data, namely, time series prediction model, neural network model based on genetic algorithm and cluster analysis.

Taking 2023 MCM C Project-Wordle as an example, by building the above three models, we predict a gradual decline in the number of reported results in the relevant games, and the number of reported results on March 1, 2023 is about 15,158, and we also get a prediction of the relative percentage of specific words, and divides the attributes of words into easy and difficult categories and so on.

As a result, we were able to better respond to the growing market and changing needs of players, and we were able to figure out if the game needed to be upgraded further, so that we could adjust the difficulty of the game, improve the rhythm of the game, etc. In this way, game quality can be improved, the user's sense of game experience can be enhanced, and the value and practical significance of statistics in game data can be realized.

Statistics plays an important role in various fields, providing people with effective methods to analyze and process all kinds of data, so that we can make more intelligent decisions, thus promoting the progress and development of society.

REFERENCES

[1]John Bridgwater."Mixing of powders and granular materials by mechanical means——A perspective." Particuology 10.04(2012):397-427. [8]Jin Hua Xia, ,and Dong Rong Liu."Research on System Safety Assessment Based on Optimized BP Network." 930.439-440(2010):469-474.

[2]C.Patrignani;K.Agashe;G.Aielli;C.Amsler;M.Antonelli;D.M.Asner;et al."Experimental Methods and Colliders." Chinese Physics C 40.10(2016):429-516.

[3]Runway Capacity and Delay under Demand VolatilityDönmez Kadir;Aydoğan Emre;Çetek Cem;Maraş Erdem Emin.The Impact of Taxiway System Development Stages on Runway Capacity and Delay under Demand Volatility[J]. Aerospace, 2022, 10(1) : 6-6.

[4]Hai-Ling Zhang;Chang-Xin Liu;Meng-Zhen Zhao;Yi Sun."Economics, fundamentals, technology, finance, speculation and geopolitics of crude oil prices: an econometric analysis and forecast based on data from 1990 to 2017." Petroleum Science 15.02(2018):432-450.

[5]Cao Bohan;Yin Qishuai;Guo Yingying;Yang Jin;Zhang Laibin;Wang Zhenquan,et al. Field data analysis and risk assessment of shallow gas hazards based on neural networks during industrial deep-water drilling[J]. Reliability Engineering and System Safety, 2023, 232

[6]Feng Hua, Zhou Fang,and Tong Qiu."Application of convolutional neural networks to large-scale naphtha pyrolysis kinetic modeling." Chinese Journal of Chemical Engineering 26.12(2018):2562-2572.

[7LIU Feng;YANG Fei;ZHAO Yu-guo;ZHANG Gan-lin;LI De-cheng. "Predicting soil depth in a large and complex area using machine learning and environmental correlations." Journal of Integrative Agriculture 21.08(2022):2422-2434.

[8]."Theory of loop algebra on multi-loop kinematic chains and its application." Science in China(Series E:Technological Sciences) .04(2007):437-447.

[9]Wenwen WANG, et al."Extracting Soil Moisture from Fengyun-3D Medium Resolution Spectral Imager-Ⅱ Imagery by Using a Deep Belief Network." Journal of Meteorological Research 34.04(2020):748-759.

[11]Zhang Jincan;Hou Xuefeng;Liu Min;Yang Shi;Liu Bo;Wang Jinchan;Zhang Juwei. Hybrid small-signal modeling of GaN HEMTs based on improved genetic algorithm[J]. Microelectronics Journal, 2022, 127.

[12] McBratney A B & De Gruijter J J A. Continuum Approach to Soil Classification by Modified Fuzzy K-means with Extragrades. [J]. Journal of Soll Science, 1992,43(1):159-175

[13] Davies D L & Bouldin D W.A. Cluster Separation Measure. [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1979(2):224-227

[14]Henghui ZHAO; Yanhui LI; Fanwei LIU; Xiaoyuan XIE; Lin CHEN. "State and tendency: an empirical study of deep learning question&answer topics on Stack Overflow." Science China(Information Sciences) 64.11(2021):131-153.

[15] Cheng Wang;Mingtao Huang;Congcong Chen;Yuancheng Li;Na Qin;Zijian Ma, et al."Identification of A-to-I RNA editing profiles and their clinical relevance in lung adenocarcinoma." Science China(Life Sciences) 65.01(2022):19-32.