Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE231061

Based on Data Mining Algorithm of Data Mining Research

Jia ZHU[1], Jun WANG
Shandong Vocational College of Commerce, Jinan 250103 Shandong China

Abstract. This paper mainly studies the data mining algorithm in big data mining, and makes improvements in view of the large amount of data, sampling noise and many errors, which can not accurately locate the data. In this paper, the basic algorithm of data mining is studied in depth, from the algorithm principle, cognitive analysis, algorithm model and so on. Neural network optimization, clustering algorithm and other combined algorithms are introduced. The parameters of standard data mining algorithm are studied to understand the application significance and function of different parameters in the algorithm. Secondly, an improved data mining algorithm is proposed to solve the defect that ordinary data mining algorithm is easy to fall into wrong data. The clustering algorithm is used to test the performance of the improved data mining algorithm, and the test results are compared with the test results of the ordinary data mining algorithm. It proves that the improved data mining algorithm has better performance in terms of data mining ability and accuracy. A new clustering data mining algorithm based on neural network is proposed. The final results show that the optimized data mining algorithm has a more perfect and intelligent data mining system.

Keywords. Data mining; Neural network; Clustering algorithm; Data processing

1. Introduction.

With the rapid development of the Internet and information technology, big data mining has become an effective means to solve practical problems such as data analysis, prediction and decision making. Big data mining technology based on data mining algorithm can carry out comprehensive and in-depth mining and analysis of massive data, obtain useful information and knowledge, and provide decision support and strategic guidance. To be specific, big data mining technology based on data mining algorithm adopts various data analysis methods, such as classification, clustering, association rule mining, anomaly detection, etc., to dig and analyze massive data in depth to discover hidden patterns, rules and relationships. It can also use machine learning algorithms that allow computers to identify and analyze data by learning from historical data sets and known knowledge to better understand the meaning of the data.

As a popular research direction in computer field, data mining algorithm is generally divided into two main research directions, namely clustering algorithm and classification

[1] Corresponding author: Jia ZHU, Shandong Vocational College of Commerce, Jinan 250103 Shandong China, E-mail: 13064092152@163.com

algorithm. First we focus on clustering algorithms. Li et al. [1] proposed a sorting clustering algorithm based on Gini coefficients of multi-reunion class solutions, and attempted to solve dynamic problems in data stream processing. Luu et al. [2] combined self-organizing mapping and convolutional neural networks to propose a new hierarchical clustering algorithm, which is suitable for time series data. Fan et al. [3] proposed a smooth L1 norm subspace clustering algorithm, which has outstanding effects in reducing sparsity and noise effects. Cao et al. [4] proposed a robust k-means algorithm aiming at the problem that the traditional K-means clustering process is easily disturbed by outliers. In addition to clustering algorithms, there are many other articles focusing on classification algorithms, which promote the mining of data sources through classification computation. Based on extreme learning machine and multi-core method, Su et al. [5] proposed a multi-core learning algorithm based on automatic kernel selection, which is suitable for classification problems under unbalanced data sets. Huang et al. [6] proposed a hybrid model algorithm for energy data prediction by combining convolutional neural network and dynamic Bayesian network. Chen et al. [8] proposed an invalid data detection and correction algorithm based on multi-rate streams. In general, these 8 papers show the hot spots and trends of current data mining algorithms, including clustering, classification algorithm in practical application of problem solutions. With the coming of the era of big data, data mining technology will face new challenges and opportunities. As a further application in the field of big data, the category of data algorithms should also get new optimization progress. Among them, Li et al. proposed a reliable and privacy-protecting speed suggestion system based on blockchain technology, which can provide drivers with real-time road condition information and help reduce the occurrence of traffic accidents. On the other hand, Mumtaz et al. [9] studied intelligent direct LTE communication and energy saving. Their proposed method can reduce the power consumption of devices and extend battery life, while improving network performance and communication quality. In addition, they also studied how to conduct resource allocation and interference management in LTE-D2D communication [10] to improve transmission efficiency and resource utilization, so as to achieve more efficient wireless communication. These studies extend our understanding of intelligent transportation and LTE communications and provide new solutions for related industries..This method combines neural network and clustering algorithm, and has a more efficient data screening mode. In the future, we will continue to improve this approach and try to apply it to other fields to provide better support for digital transformation in other industries.

Although big data mining algorithms have been widely used in industry, some problems still need to be solved, such as how to obtain valuable information from massive production data and conduct accurate analysis. In order to solve this problem, a new big data mining method based on neural network and clustering algorithm is proposed in this study, which is applied to large-scale manufacturing process monitoring. First, we use data cleaning and preprocessing methods to screen the original data, and use feature selection algorithm to determine which features have the greatest impact on the target variables and exclude other irrelevant factors. We then used clustering algorithms to categorize large datasets to better understand and visualize the data. Next, we use a supervised learning approach based on multi-layer perceptron (MLP) neural networks to construct a classifier and train the classifier to classify and predict the data. Finally, we evaluate the model, and the experimental evaluation results prove the feasibility and effectiveness of our method. In the experiment, we tested the monitoring of large data sets, including quality detection and fault diagnosis. Through continuous tuning and optimization, we can get more accurate and reliable results, and can detect more data processes and data systems. To improve the efficiency of data cleaning and selection. In conclusion, this study proposes a new data mining method, which aims to obtain valuable information from massive manufacturing process data and achieve accurate analysis and prediction.

2. Method

deep neural networks and related technologies are widely used in various industries. In the past decade, deep neural networks (DL) have achieved great success in many applications such as face recognition and machine translation. Researchers and industry experts have been trying to apply deep neural networks to more common life scenarios to solve more complex tasks and provide better results. Artificial neural network is a new hot spot in the field of artificial intelligence in the 1980s. It simulates the way the human brain processes information and builds specific computational models. Among them, a large number of neurons (or nodes) are the basis for the design of artificial neural network, and its model is shown in Figure 1. As the basic information processing unit of neural network, neurons are interconnected to form artificial neural network, and different types of networks are formed according to different connections and combinations.

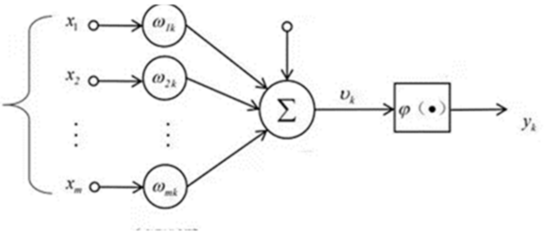

Figure.1 Neural unit assembly

Wherein, the synapse is characterized by its weight ωki to connect the two neurons, that is, the input signal xi connected to the synaptic i of neuron k is multiplied by the synaptic weight ω of k. The sum of the sum of the input signals xi weighted by the corresponding synapses of the neurons is formed by a linear combinator. The external bias bk increases or decreases the network input of the activation function accordingly, depending on whether the sum is positive or negative. The activation function φ, as a nonlinear transformation, is used to limit the output amplitude of neurons, that is, to limit the output signal within a small interval (usually the normal amplitude range of a neuron's output can be written as the unit closed interval [0,1] or another interval [-1,+1]). The model of neuron k described by mathematical formula is shown in Formula 1:

$$y_k = \varphi(\sum_{i=1}^{m} \omega_{ik} \, x_i + b_k) \tag{1}$$

Indeed, the increasing data volume, model size, and output accuracy have led to the successful application and mass popularization of deep neural networks. The characteristics of deep neural network can be mainly attributed to nonlinear transformation. Conventional methods such as matrix factorization, decomposition machines, sparse linear models, etc. are linear models in nature, and the linear assumptions underlying these traditional models are often too simple and greatly limit the expressiveness of their modeling. In contrast to linear models, deep neural networks can use activation functions such as sigmoid, tanh, and relu for nonlinear modeling of data, as shown in Figure 2.

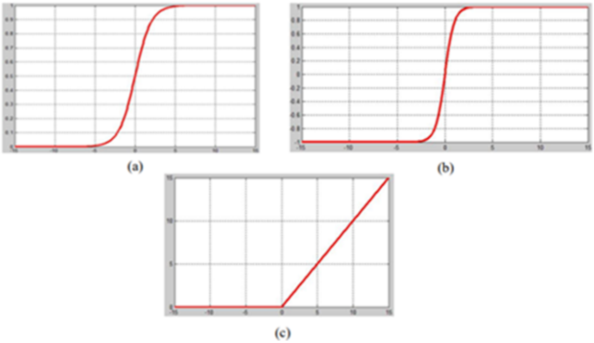

Figure 2. Activation function of deep neural network

In this paper, using the expectation maximization algorithm clustering algorithm (EM). EM algorithm is named because it is an iterative optimization process, each iteration is divided into two steps, one of which is the expected step (E step), the other is the maximum step (M step). The basic idea is: firstly, the values of model parameters are estimated according to the given observational data; Then, the value of missing data is estimated according to the parameter value estimated in the previous step, and the parameter value is estimated again according to the estimated missing data and previously observed data, and then iterated repeatedly until finally convergence, the end of iteration. The derivation of EM algorithm relies on two basic knowledge: "maximum likelihood estimation" and "Jensen's inequality". Maximum likelihood estimation is an application of probability theory in statistics. It is a method of parameter estimation. When a random sample is known to satisfy a certain probability distribution, but the specific parameters are not clear, the parameter estimation will observe the results of several tests, and use the results to derive the approximate value of the parameters. If the sample set 12 x the probability density of p (x),the joint probability of sample set as:

$$L(\theta) = L(x\_1, x\_2 \cdots, x\_n; \theta) = p(x;\theta), \theta \in \delta \tag{2}$$

Reflects the probability when the parameters of the probability density function is i, get the probability of this set of samples. When the largest probability, i.e., likelihood function L(i) is the largest, then called θ maximum likelihood estimator, remember to:

$$\theta = argmaxl(\theta) \tag{3}$$

Maximizing the likelihood function L (i), then the maximum corresponding θ is required parameter values. When you maximize the function, you take the derivative of it, and then you set the derivative to zero and you solve the equation. If the i contains multiple parameter vector, all parameters of partial derivative, then n unknown parameters, and then solve the n equations, equations is an extreme value of the likelihood function, and the n parameters is obtained. Derivation process, in order to simplify the function will take logarithm L (i), define the logarithm likelihood function, remember to:

$$H(\theta) = lnL(\theta) = \sum_{i=1}^{n} lnp(x_1\,; \theta) \tag{4}$$

Jensen's inequality is defined as: Let f be a function whose domain is real, and if for all real numbers X, the second derivative of f(X) is greater than or equal to 0, then f is convex. When X is a vector, if its hessian matrix H is semi-positive definite, then f is convex. If it's only greater than 0, not equal to 0, then f is strictly convex. In particular, if f is strictly convex, if and only if X is constant, Jensen takes the equal sign and Jensen's inequality is stated as:

$$E[f(X)] \geq f(EX) \tag{5}$$

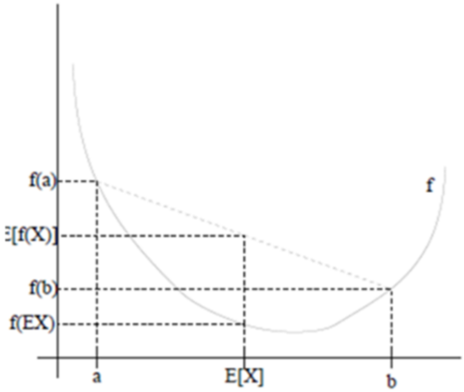

Figure 3. Iterative particle swarm verification

The pattern can be seen in figure 3. The maximum likelihood estimation mentioned above is to solve the parameters of a given random sample distribution. If a sample set is encountered and the corresponding category of each sample is unknown, the EM algorithm should be used. For the training samples that are independent of each other among the samples, the existence of class z corresponding to each sample i is unknown, which is also called implicit variable. At this moment need to estimate the parameters of the probability model p (x, z), by maximum likelihood estimate the evolution of the formula is as follows:

$$\sum_i lnp\left(x^{\{i\}};\theta\right) = \sum_i ln \sum_x p(x,z) \tag{6}$$

After the first equal sign, class variable i is added on the basis of the original maximum likelihood estimation formula. For all possible class z of each sample i, the sum of the joint probability density function on the right side of the equation is obtained, and the likelihood function of random variable x is obtained on the left side of the equation. However, since the logarithm of the sum is very complicated, the second equal sign and the third equal sign are converted into the formula, and Jensen's inequality is used to convert the logarithm of the sum into the sum of the logarithms, greatly simplifying the calculation difficulty. On the evolution of the type can be seen as the likelihood function L(i) lower boundary and its maximum is inequality into equation, based on Jensen inequality, want equation was set up, need to make a random variable constant.

$$\frac{p\left(x^{(i)},z^{(1)},\theta\right)}{Q_I(z^i)} = c \tag{7}$$

So far, after fixing other parameters, the formula for calculating Q is the posterior probability, and step E is to solve the problem of how Q chooses and establish the lower bound of l. The M step is to adjust i after i is determined, and then continuously adjust l larger. As the lower bound increases, the maximum likelihood estimator monotonically increases and thus converges. EM algorithm steps is first initialized θ distribution parameters, then repeat steps and M E steps until convergence. Based on the initial value of the parameter or the model parameters of the last iteration, the posterior probability of the implicit variable, namely the expectation of the implicit variable, is calculated as the current estimated value of the implicit variable, which is described as follows:

$$Q_I\left(z^i\right) = p(z^i|x^i, \theta) \tag{8}$$

3. Experiment

The functions of data mining usually include four parts: characterization and differentiation, frequent pattern mining, classification and regression, clustering and outlier analysis. Class/concept description refers to describing each class and concept in a summary, concise and accurate way, which is usually realized by data characterization and differentiation, as shown in Figure 4. Frequent patterns are those that occur frequently in data. Mining frequent patterns can reveal interesting associations hidden in the data. Frequent item set mining is the basis of frequent pattern mining, and related theories and algorithms will be described in detail in Chapter 4. A class of data mining tasks for predictive analysis includes classification and regression. Where, the derivation of the classification model is based on the analysis of the training data set (data objects with all class labels known) and the prediction of class labels of objects with unknown class labels. Examples of various export models are shown in Figure 4. The research and application of neural network algorithm in classification task will be described in detail in Chapter 3 and 4. In addition, regression analysis, as one of the most commonly used numerical prediction statistical methods, is used as the guarantee means of data mining in this paper, so as to improve the accuracy of data mining.

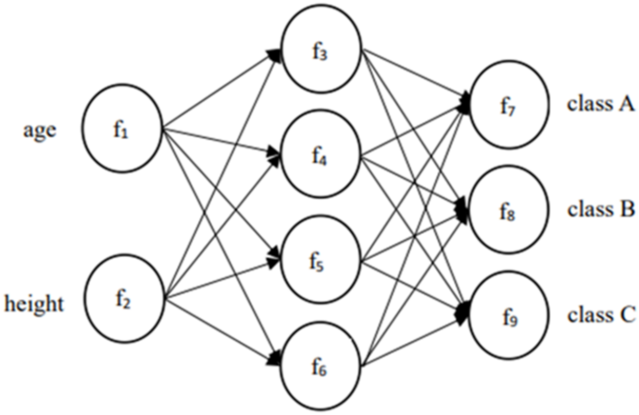

Figure 4. Data mining classification model

Considering the continuous increase of data capacity, the wide distribution of data and the computational complexity of some data mining algorithms, it is necessary to ensure that the running time of mining algorithms in processing massive data is within the acceptable range, and the spatial complexity is not beyond the range of running machines. One solution is to develop parallel and distributed data-intensive mining algorithms. This algorithm first divides the data into several "fragments", then processes each fragment in parallel, and finally combines the patterns or rules explored by each part to discover knowledge. Cloud computing and cluster computing use distributed and cooperative computers to handle large scale computing tasks, which is an important direction of parallel data mining. In addition, the development of incremental data mining is driven by costly mining processes and incremental inputs, which incrementally iterate, modify, and enhance the industry's discovered knowledge by binding with new data updates. In a word, the key indicators to evaluate mining algorithms need to take into account (1) effectiveness: the running time is predictable and acceptable; (2) Scalability: The model or rule has certain adaptability to the increase or decrease of data volume and the change of data type. This algorithm belongs to the use of clustering algorithm in adaptability. Finally, the clustering algorithm is used to draw experimental conclusions. Compared with ordinary data mining methods, the accuracy of clustering data mining algorithm is greatly improved. Especially in the face of error data processing and big data, pre-processed data has a stronger advantage.

The success of this model lies in the change from a single model dealing with multiple classification problems to multiple models dealing with several binary classification problems. On the premise of ensuring accuracy, the internal structure complexity of neural network model is reduced effectively, and the network model of structural interpretability is constructed, which is conducive to the further study of interpretability theory. Another reason for the effectiveness of the model is that fastText, which has a significant speed advantage in text classification tasks, is selected as the basic model. If traditional models such as RNN and LSTM are used to embed NNF framework, the time cost will be huge under the premise of limited computing resources in the laboratory, so it is not feasible to choose NNF scheme at this time.

4. Discussion

Big data mining is a data-based process to discover the value and insights hidden behind big data by applying techniques such as machine learning and statistical analysis. Its research scope includes data processing, feature selection, model construction, result interpretation and so on. In the study of big data mining, it is necessary to define the research objectives and problems, and determine the methods and ways of collecting, cleaning and sorting data. Subsequently, data preprocessing, such as feature extraction, dimension reduction and normalization, can be carried out to facilitate subsequent model training and prediction. Then, an appropriate machine learning or deep learning algorithm should be selected, and the model should be evaluated and optimized by means of cross-validation and parameter adjustment. After a good model is developed, the generalization ability and performance can be evaluated using test data sets, and valuable information and association rules can be mined from them by combining visualization tools and interpretive analysis methods. Finally, the mining results are combined with actual business scenarios to provide relevant decision support and action suggestions, and the model is constantly fed back and revised to achieve continuous data-driven growth and innovation.

Big data mining research aims to dig out the value and potential contained in big data by means of data science, so as to provide strong support for realizing commercial value and social development. In order to achieve better data mining, more advanced machine learning models, such as deep neural networks, can be considered and analyzed in combination with actual market risks and political factors to improve the accuracy and reliability of the forecast.

5. Conclusion

Data mining is a powerful tool that can help people dig out the underlying patterns and patterns in massive amounts of data. In the field of big data algorithm mining, this paper uses the principle of combination algorithm. Neural network is one of the commonly used algorithms and plays a very important role in data mining. Neural networks can train data and discover complex nonlinear relationships through learning. In the face of a large number of data mining work, the use of more advanced data mining algorithm model can better improve the efficiency. Another common algorithm is clustering, which groups similar data points together so that we can better understand the logic behind the model. Clustering algorithms can not only identify outliers, but also divide massive data sets into groups to analyze the data in more detail. By combining advanced technologies such as neural network and clustering algorithm, data mining can provide more accurate, comprehensive, timely and scientific data analysis results, provide better decision support for enterprise personnel, and reduce risks and losses.

References

[1] Yanhua Li, Jian Li, Zhihao Xie. Ranking-based clustering with multiple clustering solutions in the context of data streams. *Information Sciences*, vol. 523, pp. 20-32, 2020.

[2] Thanh Luu, Shuo Zhang, Shengrui Wang, et al. Hierarchical clustering for time-series data based on self-organizing maps and convolutional neural networks. *Information Sciences*, vol. 514, pp. 115-127, 2020.

[3] Xinxin Fan, Yichi Zhang, Zhong Ma, et al. Smooth l1-norm subspace clustering. *Neurocomputing*, vol. 385, pp. 226-235, 2020.

[4] Jianting Cao, Lu Wan, Ning Zheng, et al. Robust k-means algorithm using worst-case deviation difference. *Soft Computing*, vol. 23, pp. 2635-2645, 2019.

[5] Hongli Su, Huan Liu, Chao Cheng. Multi-kernel extreme learning machine with automatic kernel selection for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2791-2803, 2019.

[6] Chaohao Huang, Yongquan Zhou, Yuping Qin, et al. A hybrid model based on convolutional neural network and dynamic Bayesian network for predicting energy consumption. *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 4896-4905, 2019.

[7] Xuanyang Xi, Jie Li, Daoliang Li, et al. A novel ensemble learning algorithm for imbalanced intrusion detection datasets. *Information Sciences*, vol. 481, pp. 79-93, 2019.

[8] Xiaoyang Chen, Jingtao Yao, Senzhang Wang. Detection and correction of invalid data in multi-rate stream. *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 7, pp. 1351-1363, 2019.

[9] S. Mumtaz, H. Lundqvist, K. M. S. Huq, et al., "Smart Direct-LTE Communication: An Energy Saving Perspective," *Ad Hoc Networks*, vol. 13, pp. 296-311, Mar. 2014.

[10] S. Mumtaz, K. M. S. Huq, A. Radwan, et al., "Energy Efficient Interference-Aware Resource Allocation in LTE-D2D Communication," in Proceedings of the 2014 *IEEE International Conference on Communications (ICC)*, Sydney, NSW, Australia, 2014..