Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2023.040108
ISSN 2790-0983 Vol. 4, Issue 1: 61-69

Optimization Algorithm of Big Data Mining Based on Machine Learning Model

Yongfeng Shi*

China Water & Power Press, Beijing 100043, China

shiyongfeng@mwr.gov.cn

**corresponding author*

Keywords: Machine Learning, Big Data, Data Mining, Optimization Algorithm

Abstract: With the arrival of the big data era, massive data storage, massive information push, and massive complex products and services are flooding the whole society. As an important and effective method, machine learning model has been widely used in various fields. In order to apply data mining technology to practice well, this paper makes an in-depth study on machine learning model and data mining optimization methods. This paper mainly uses the methods of experiment and comparison, and puts forward the advantages of various improved algorithms through the detection of fragments, orchids and beverages. The experimental results show that the C index obtained by ESPSO-FCM on fragments is 0.523, which is larger than the other two algorithms. ESPSO-FCM algorithm is an improved clustering algorithm with higher convergence accuracy, stronger partition ability and better clustering quality.

1. Introduction

In our life, big data mining is a topic worthy of research and exploration[1-2]. Based on the optimization algorithm of machine learning model, through the processing of the effective information stored in the existing data, the relevant formulas and algorithms are used to realize big data mining[3-4]. This method can effectively solve the problem of low accuracy due to collinearity and random errors in traditional large sample problems[5].

This paper studies big data optimization algorithm *Shao Z* An optimized mining algorithm based on dynamic data analysis of students' learning degree is proposed. The algorithm first uses the optimized text classification technology to automatically match the problem text to the knowledge points, thus improving the efficiency and quality. Then, based on the dynamic data of students' response records, the subjective weighting method and expert experience are used to

generate the learning degree matrix of students on knowledge points. Finally, DBSCAN clustering algorithm is used to cluster students' personalized learning characteristics according to the learning degree matrix[6].*Behravan I,M*This paper studies an automated big data clustering method based on swarm intelligence, which automatically clusters the behavior core of players in football, and extracts various roles in football. This paper uses PSO algorithm to establish a new solution. First, the method looks for the number of clusters in the solution process, and then determines the location of the cluster center in the solution process. Through six complete experimental platforms, the correctness of the algorithm is verified[7].*Zhang S*The importance of microarchitecture events of big data tasks is ranked, and the performance big data dimension is reduced to optimize the big data algorithm according to the described performance characteristics. There is no doubt that the comprehensive monitoring of sports training process is a complex system engineering. The main monitoring includes three aspects: physical condition, technical and tactical skills and intelligence. Sports technology is embodied in[8].

This paper first studies the machine learning model, and discusses the machine learning and its model construction[9-10]. Secondly, the related algorithms of data mining are improved, and the improved APRIORI algorithm is obtained[11-12]. Then, the system of data mining is constructed and its algorithm optimization is discussed. Finally, through the way of experiment, three kinds of objects are tested on the spot and relevant data results are obtained[13-14].

2. Optimization Algorithm of Big Data Mining Based on Machine Learning Model

2.1. Machine Learning Model

At present, the main machine learning models can be predictive. Of course, models can also be both. Therefore, the main object involved in machine learning is learning algorithm. Machine learning frameworks mainly include Hadoop's Mahout, Spark's MLLb and Graphlab. The machine learning model is mainly based on Spark Core, on which KMeans clustering model, ALS based collaborative filtering recommendation model and Spark Streaming based online KMeans clustering model are implemented. As for the invocation interface of the model, it is designed as a Web Service interface. In the experiment, it is mainly in the form of direct invocation.

In the learning process, a large amount of data will be generated. What we have studied is all data sets at a point and in a time period. We need to store these discrete and non real time information together in a continuous form. Machine learning model is a purposeful data analysis method[15-16]. It uses the knowledge that has been mined from the existing computer to implement new rules by operating the existing database. In the massive information of big data, we can optimize the results obtained through machine learning model. At present, the commonly used algorithms mainly include graph based method, statistical analysis method and artificial neural network. In big data, all our information is stored in it and what we need to do is to analyze and process these massive data and better mine useful and effective information. In the learning process, artificial neural networks often fall into local optimization due to their large training samples, which cannot ensure the accuracy and effectiveness of big data analysis results[17-18].

2.2. Improved APRIORI Algorithm

Association rules can easily find some interesting relationships or associations between different WEB datasets. In the case of the potential relationship between customers' purchase of goods, an interesting phenomenon can be discovered by using association rules, that is, some general rules of

customers' purchase behavior. In order to find all the largest item sets in the massive data using APRIORI algorithm, there is a necessary way to connect and operate layer by layer. The operation process is mainly through continuous scanning of the database. Because a large number of candidate item sets will be generated in the calculation of each layer, it is necessary to compare each candidate item set with the transaction records stored in each database. In this way, the cycle scanning will continue until no new candidate item sets are generated, and finally all frequent item sets that meet the latest support threshold will be obtained. The second step is to mine all the association rules of the data through the found frequent itemsets. It is not difficult to find that the APRIORI algorithm requires a lot of storage space and time to find all frequent itemsets, which is the essential reason why the APRIORI algorithm is inefficient. Apiori mode information set retrieval frequency is too high, statistical effect is poor, and data storage and query efficiency is very low. Therefore, the algorithm needs to be improved.

APRIORI can be improved from the following points: In order to save memory consumption and improve I/O performance, it is undoubtedly a wise choice to reduce the number of database scans. The frequent itemsets are generated circularly and progressively, but many of them are redundant. Therefore, the computational performance of the process of generating frequent itemsets can be improved. Starting from the essence of APRIORI, research on more convenient derivation of this type of operation. The mechanism of random target call is introduced to reduce the computational complexity so as to greatly simplify the working efficiency of the system. In order to improve the storage and query efficiency of the algorithm, we can consider changing the original storage structure.

Only scanning the entire database can determine whether a set is frequent or infrequent in all candidate sets. Because this step is time-consuming and laborious, if you can eliminate some infrequent item sets in advance, you can greatly save the scanning workload and improve the algorithm efficiency. In addition, if the number of elements in the candidate set can be minimized, the computation can also be saved. Under the guidance of this idea, if the two methods can be combined, that is, after obtaining k-dimensional frequent item sets, remove some non frequent item sets, and organize them to be combined into candidates again to generate redundant generation, then the generation of (k+1) dimensional candidate sets will be greatly simplified.

2.3. Data Mining

Big data mining refers to extracting valuable information and potentially useful effective structures from a large number of web pages. In the traditional database retrieval process, it is often necessary to find target attributes or keywords to obtain relevant links. Mining model building method this technology is a machine learning algorithm based on artificial neural network. It can meet people's needs and bring benefits, and even change the new methods of various business models in the existing market environment. The core of data mining is big data processing. Under the traditional mode, how to analyze and effectively use a large amount of information has become a problem. Based on the machine learning model, a model containing multiple eigenvalues can simply describe the complex relationship in multiple dimensions. In the analysis of big data, we usually need to find the most useful and satisfactory results after processing the information.].

Classification is a very important task in data mining. Clustering is to divide data into different data classes according to their different characteristics. There are many algorithms for data mining, and we need to decide which algorithm to use according to the specific situation and application requirements.

To sum up, the system of data mining algorithms can be shown in Figure 1:

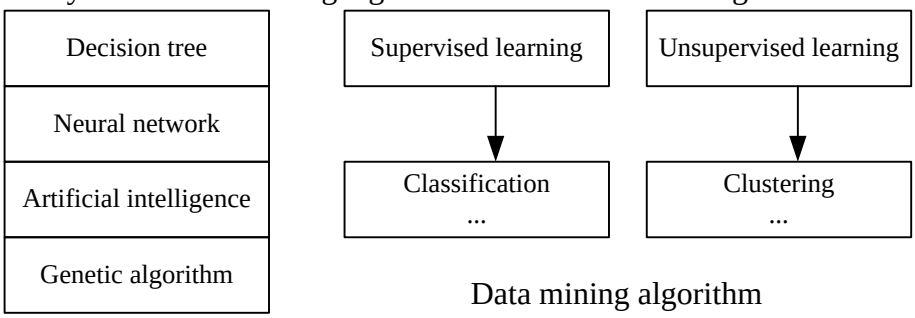

Figure 1. The system of data mining algorithm

Traditional data mining is the basis of Web data mining. Therefore, traditional data mining and Web data mining have similarities in the process. However, due to the characteristics of Web data mining, there are differences in the specific mining process. Typical Web data mining includes four steps: data collection, data preprocessing, pattern discovery and pattern analysis. Web usage mining is to mine effective information according to the traces left by users on the Internet.

PageRank not only focuses on the direction between pages, but also takes the authority of the page as the evaluation standard. By combining the number of links and the quality of web pages, a more standardized web page evaluation standard can be obtained. The value obtained by PageRank operation is an important indicator used to evaluate the page, which is not related to the keywords entered by visitors, that is, the algorithm is not related to the keywords. If there is a retrieval system, its calculation will not take the page content as an evaluation index. The information obtained by this retrieval system is the same for any different keyword, that is the page with the highest PageRank value.

FCM algorithm is the most widely used fuzzy data mining algorithm, but the algorithm is easy to fall into local extremum, which requires improvement of FCM algorithm. In order to overcome the shortcomings of FCM algorithm, an adaptive particle swarm optimization algorithm is introduced. For all test functions, ESPSO algorithm has better convergence and higher accuracy than SPSO algorithm. The ESPSO algorithm is closer to the global optimal solution, which shows that the improved algorithm in this chapter has better convergence and higher accuracy.

3. Simulation Experiment

3.1. Experimental Preparation

The experimental platform of this paper is Windows 13 system and the algorithm is implemented using MATLAB R2016a software. This paper tests the data set in UCI machine learning database, and selects FCM, FCM-PSO and ESPSO-FCM algorithms to compare the clustering results of the data set.

3.2. Experimental Setup

The parameters of the algorithm used in the experiment are set as follows: the inertia weight of FCM-PSO algorithm decreases linearly from 9/10 to 1/10 with the iteration. In the ESPSO-FCM algorithm, the inertia weight decreases linearly from 9/10 to 3/10 with the iteration. For the sake of fairness, the fuzzy index of all algorithms is 3, and the population size of FCM-PSO and

ESPSO-FCM is set to 20. When the FCM reaches 30 iterations or the change of the target function Jm is less than or equal to 0.0001, the operation is terminated; when the FCM-PSO and ESPSO-FCM reach 50 iterations or the change of the target function Jm is less than or equal to 0.0001, the operation is terminated.

3.3. Experimental Scheme

In this paper, three sets of standard test data sets were used for the experiment: orchids, beverages and fragments. In order to accurately understand and analyze the performance and clustering effect of ESPSO-FCM algorithm, two groups of tests are carried out in this part. The first group of experiments evaluates the quality of the clustering results of the algorithm through typical clustering effectiveness indicators to test the feasibility and effectiveness of the algorithm. The second group of experiments used the objective function of fuzzy clustering (Jm) as the evaluation standard to test the accuracy of the algorithm clustering. The following four typical cluster effectiveness indicators are selected:

The larger the value of partition coefficient C, the better the clustering result. Represented by Formula (1):

$$C = \frac{1}{\mathrm{m}} \sum_{i=1}^{x} \sum_{k=1}^{m} v_{ik}^{2} \tag{1}$$

The smaller the partition entropy E is, the better the clustering result is. The formula is shown in (2):

$$C = \frac{1}{\mathrm{m}} \sum_{i=1}^{x} \sum_{k=1}^{m} v_{ik} \log_{a}(v_{ik}) \tag{2}$$

The other two indicators are typical for measuring intra class compactness and inter class dispersion. The smaller the X index, the better the clustering result. Small P index means that the intra class is not compact or the inter class is not discrete during classification. On the contrary, a large P index indicates good intra class compactness and inter class dispersion of clustering. The larger the P index is, the better the clustering is divided.

4. Experimental Results and Analysis

4.1. Comparison of Clustering Results

In this experiment, four fuzzy indexes C, E, B, P are used to verify the feasibility and effectiveness of ESPSO-FCM algorithm. The three algorithms FCM, FCM-PSO and ESPSO-FCM are compared on all datasets. The average and standard deviation of C and E index values in the effectiveness of fuzzy clustering are shown in Table 1:

Table 1. Comparison of clustering results about the C indicators and the E indicators

	C			E		
	FCM	FCM–PSO	ESPSO–FCM	FCM	FCM–PSO	ESPSO–FCM
Orchid	0.783	0.775	0.783	0.396	0.412	0.396
Drink	0.791	0.789	0.791	0.384	0.386	0.381
Patch	0.502	0.495	0.523	0.969	1.09	0.925

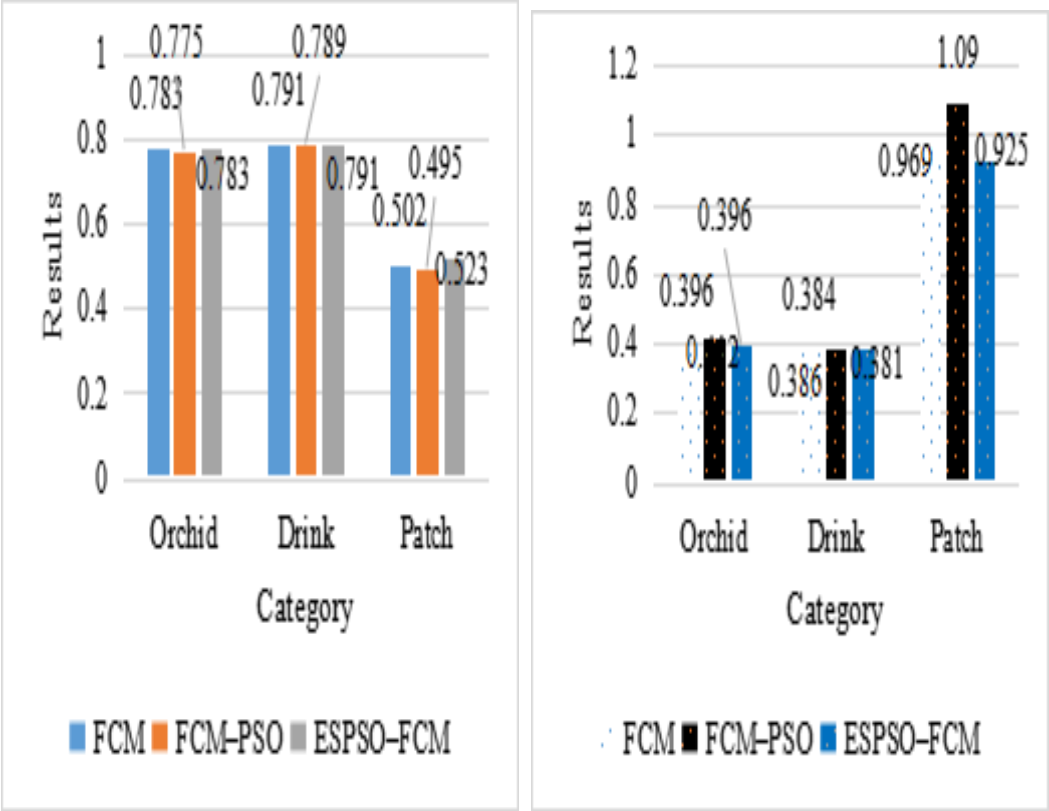

Figure 2. Comparison of clustering results about the C indicators and the E indicators

As shown in Figure 2, we can see that FCM and ESPSO-FCM perform equally and better than FCM-PSO for orchid and beverage datasets. However, it performs better than FCM on more complex data set fragments. ESPSO-FCM obtains the smallest E index on all data sets.

Table 2 shows the comparison of clustering results of FCM, FCM-PSO and ESPSO-FCM algorithms on X index and P index. The results of orchid and beverage data sets are similar. Although FCM-PSO has the smallest X index and the best performance in the fragment data set, the overall performance of ESPSO-FCM algorithm is stable.

Table 2. Comparison of clustering results about the X indicators and the P indicators

	X			P		
	FCM	FCM–PSO	ESPSO–FCM	FCM	FCM–PSO	ESPSO–FCM
Orchid	0.138	0.141	0.138	1.972	2.038	2.058
Drink	0.127	0.128	0.127	2.207	2.173	2.248
Patch	0.949	0.32	0.931	8.069	9.047	9.605

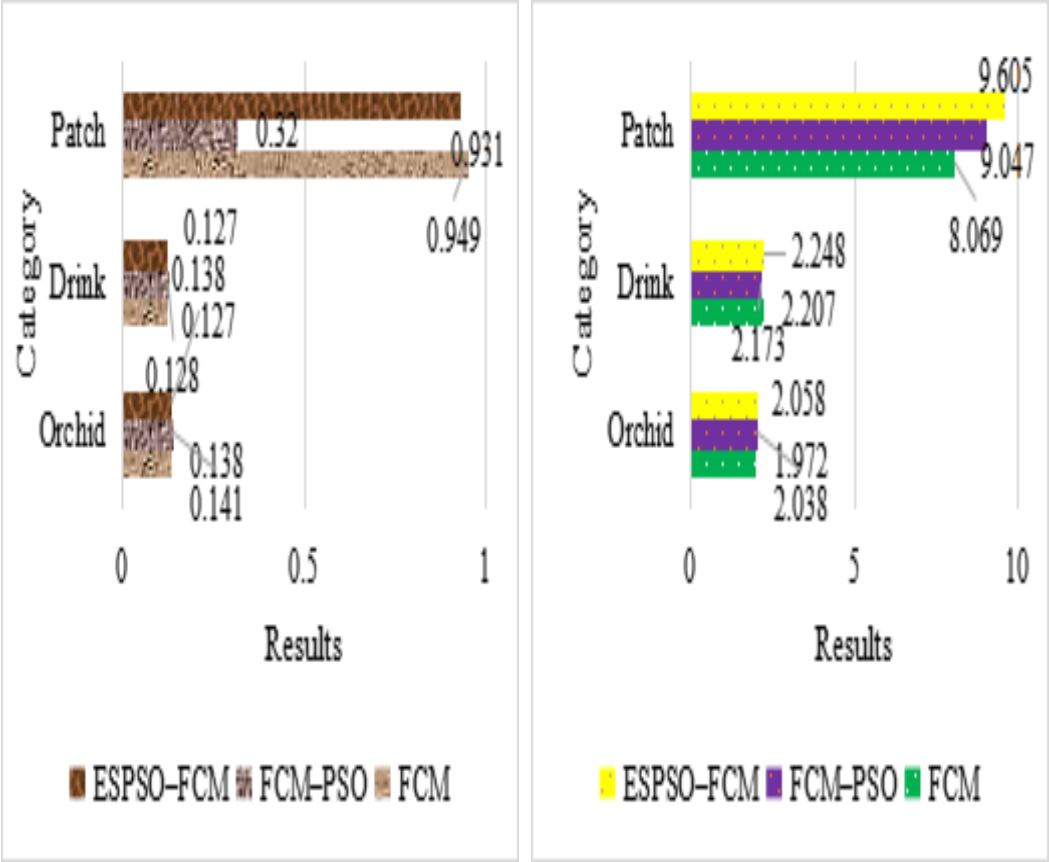

Figure 3. Comparison of clustering results about the X indicators and the P indicators

As shown in Figure 3, we can find that the FCM-PSO corresponds to a larger P index than the FCM, except for the beverage dataset. ESPSO-FCM obtained the maximum on all datasets with the best performance. As is illustrated by the results in the table, in terms of class algorithm effectiveness, the overall ESPSO-FCM algorithm has achieved the best performance steadily.

5. Conclusion

By studying traditional data mining algorithms and combining big data optimization technology based on machine learning model, this paper proposes a new form of method to solve the above problems. In this paper, we mainly use machine learning model and improved data mining algorithm to model and analyze the relevant data mentioned by the experimental object. The experimental results show that the improved algorithm can guarantee the accuracy of data processing. However, machine learning is a very complex and large span problem. Therefore, we need further research to better improve the neural network and big data processing capabilities.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Shao Z, Sun H, Wang X, et al. An Optimization Mining Algorithm for Analyzing Students' Learning Degree Based on Dynamic Data. IEEE Access, 2020, PP(99):1-1.

[2] Behravan I, Zahiri S H, Razavi S M, et al. Finding Roles of Players in Football Using Automatic Particle Swarm Optimization-Clustering Algorithm. Big Data, 2019, 7(1):35-56.

[3] Zhang S, Mao H. Optimization Analysis of Tennis Players' Physical Fitness Index Based on Data Mining and Mobile Computing. Wireless Communications and Mobile Computing, 2021, 2021(11):1-11.

[4] Yu H. Apriori algorithm optimization based on Spark platform under big data. Microprocessors and Microsystems, 2021, 80(11):103528.

[5] Yang F, Liao X. An Optimized Sanitization Approach for Minable Data Publication. Big Data Mining and Analytics, 2022, 5(3):257-269.

[6] Gaye B, Zhang D, Wulamu A. Improvement of Support Vector Machine Algorithm in Big Data Background. Mathematical Problems in Engineering, 2021, 2021(1):1-9.

[7] Wang C, Li J, Rao H, et al. Multi-objective grasshopper optimization algorithm based on multi-group and co-evolution.. Mathematical biosciences and engineering : MBE, 2021, 18(3):2527-2561.

[8]Yiqi, Wang, Yipin, et al. Model Training Task Scheduling Algorithm Based on Greedy-Genetic Algorithm for Big-Data Mining. Journal of Physics: Conference Series, 2019, 1168(3):32057-32057.

[9] Yang X, Yang J, Yang Y, et al. Data-mining and atmospheric corrosion resistance evaluation of Sn- and Sb-additional low alloy steel based on big data technology. International Journal of Minerals, Metallurgy and Materials, 2022, 29(4):825-835.

[10] Shi X, Liu Y. Sample Contribution Pattern Based Big Data Mining Optimization Algorithms. IEEE Access, 2021, PP(99):1-1.

[11] Watada J, Roy A, Vasant P. Preference Identification Based on Big Data Mining for Customer Responsibility Management. International Journal of Intelligent Technologies and Applied Statistics, 2020, 13(1):1-24.

[12] Yue H, Liao H, Li D, et al. Enterprise Financial Risk Management Using Information Fusion Technology and Big Data Mining. Wireless Communications and Mobile Computing, 2021, 2021(1):1-13.

[13] Ma H. Enterprise human resource management based on big data mining technology of internet of things. Journal of Intelligent and Fuzzy Systems, 2021(1):1-7.

[14] Du Y, Zhao T. Network Teaching Technology Based on Big Data Mining and Information Fusion. Security and Communication Networks, 2021, 2021(9):1-9.

[15] Guo L, Wang M, Lin Y. Electromagnetic Environment Portrait Based on Big Data Mining. Wireless Communications and Mobile Computing, 2021, 2021(3):1-13.

[16] Xie C, Xiao X, Hassan D K. Data mining and application of social e-commerce users based on big data of internet of things. Journal of Intelligent and Fuzzy Systems, 2020, 39(1):1-11.

[17] Suresh K, Karthik S, Hanumanthappa M. Design an efficient disease monitoring system for paddy leaves based on big data mining. Inteligencia Artificial Revista Iberoamericana de Inteligencia Artificial, 2020, 23(65):86-99.

[18] Wang L. Improving the performance of precision poverty alleviation based on big data mining and machine learning. Journal of Intelligent and Fuzzy Systems, 2020, 40(4):1-12.