Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2023.040101
ISSN 2790-0983 Vol. 4, Issue 1: 1-8

Practice and Application of Fusion Machine Learning in Data Analysis

Sujuan Han*

Qinghai Normal University, Qinghai, China

**corresponding author*

Keywords: Machine Learning, Data Analysis, Data Mining, System Design

Abstract: Machine learning is a process in which computer is used to train and calculate input data and output results in a complex, multi task simulation. In data analysis, we can use machine learning to carry out experimental research and theoretical verification. In order to improve the ability of data analysis, we need to use machine learning and data mining methods to better process data. In this paper, experimental method and principal component analysis method are mainly used to test and discuss the fusion of machine learning in data analysis. The experimental results show that the CPU utilization rate in Scheme 4 is about 85% on average. The reason why the CPU of the Scribe center server is reduced is that after receiving data, there is less data to decompress, which reduces the CPU utilization.

1. Introduction

With the development of network, all kinds of massive data need to be processed and applied. Machine learning has the characteristics of timeliness, speed and large capacity in data processing. Data mining uses machine learning to understand the useful or useless information hidden in a large amount of data. How to get the useful information we need quickly and effectively becomes more and more important. Data mining technology plays an important role in information classification and calculation. It can extract and utilize the unstructured data that cannot be directly obtained, processed and used by humans.

There are different theoretical achievements in machine learning and data analysis from many aspects. For example, some experts point out that machine learning enables machines to learn rules from various data to classify and recognize new samples. [1-2] Some experts also designed and implemented a LIS data analysis system based on web data mining and concurrent access control technology [3-4]. In addition, some experts believe that data warehouse and data mining technology

can better organize big data, make it play its best role, and provide a basis for business decision-making [5-6]. Therefore, the research on the application of machine learning in data analysis in this paper has theoretical basis and is conducive to the design of data processing and analysis system.

In this paper, deep learning and principal component analysis are first studied, and the calculation process is described. Secondly, the event category recognition of data and knowledge fusion is described. Then, statistical learning and support vector machine are discussed briefly. Finally, through the design and experiment of the data analysis system, relevant data and conclusions are obtained.

2. Practice of Fusion Machine Learning in Data Analysis

2.1. Deep Learning and Principal Component Analysis

Before data analysis, we need to preprocess the results obtained. This step is very important and complex. Generally, there are three main methods, sample classification, feature extraction and normalization. In the process of data preprocessing, we need to conduct a series of statistical analysis on the samples obtained. This will ensure that our final results are more accurate [7-8].

Deep learning has more hidden layers than traditional neural networks. In deep learning, the flat learning method greatly reduces the learning efficiency. Second, solutions to dimensional disasters. Multi level extraction and network weight storage form transformation are used to extract and transform the original data to obtain high-dimensional abstract features. Unmanned learning generates low dimensional features, providing more concise model features for subsequent processing.

The deep learning method has the following problems: First, it needs a lot of data. Secondly, the distribution of test and drive components should be consistent. It matches the current distribution. If different, it is difficult to obtain satisfactory results. Third, deep learning is slow. Fourth, parameter setting is difficult. Parameter design methods are different for different problems [9-10].

Finite boltzmann machine is a model based on energy minimization. The energy function consists of hidden layer, visible layer and weighted energy. State probability (W, K) is determined by the energy state of the whole network:

$$G(\mathrm{w},k)=\frac{1}{Q}f^{-S(\mathrm{w},k)} \tag{1}$$

Among them, $Q=\sum f^{-S(\mathrm{w},k)}$.

Take the variable in the middle:

$$\mu(a)=-\log\sum_{g} f^{-S(a,k)} \tag{2}$$

The DBN training algorithm is obtained by greedy stacking RBM layer by layer. The whole DBN network can be regarded as a deep BP network. Back propagation will spread the error information from top to bottom, so as to fine tune the whole DBN weight and solve the problem of long training time.

2.2. Identification of Event Categories Integrating Data and Knowledge

Data based event extraction is to obtain the connotation of events from corpora through machine learning, while knowledge-based event extraction method is to summarize the connotation of event

patterns or events through language using its own knowledge. The event connotation obtained by the event extraction method based on data is realized by establishing a machine learning model, which is an implicit representation with poor interpretability [11-12].

The essence of event mining process is analyzed, and it is found that event mining based on knowledge is the same as event mining based on data. The essence of the two event extraction methods is the same, which provides a prerequisite for merging the two methods into one event extraction method. And these two methods have their own advantages and disadvantages in the event extraction process. The method of merging data and knowledge can combine the advantages of these two methods, weaken the influence of their shortcomings, and improve the effect of event extraction. This paper proposes an event extraction strategy integrating data and knowledge, and designs a complete event extraction method, which describes the process of event extraction method with the essential connotation and extension of events [13-14].

This paper proposes an event extraction method that integrates data and knowledge. Event extraction includes event category recognition and event element recognition. Firstly, the event category recognition of data and knowledge fusion is studied. In event category recognition, we need to build relevant knowledge and select appropriate event features from the training corpus. Add the constructed knowledge to the event feature to form a new event feature. By processing new features, feature vectors are generated. The machine learning model is established after the feature vector is processed by the machine learning method, and the event category recognition of the test corpus is realized [15-16].

2.3. Statistical Learning and Support Vector Machine

Problems in multidimensional data analysis: In multimodal decomposition and clustering combination, due to some shortcomings in the algorithm, the program runs longer, and requires re threshold segmentation. Therefore, when a training set is processed, multiple meshes can be obtained. However, this cannot guarantee that every sample can be classified into the same area. Analyzing the correlation in the test data results will increase the calculation amount and cost for similar phenomena in different grid domains. Multidimensional data analysis algorithm is based on different methods, such as clustering technology, classification and association rules. These methods have their own uniqueness. Multidimensional data analysis and application is a dynamic process. It has been developing with time, space and technology. The multidimensional data analysis algorithm is improved on the traditional machine learning method, and other related technologies such as neural network, fuzzy clustering and support vector machine are introduced [17-18].

Support vector machine linear classifier is a basic statistical learning model algorithm. In this case, the problem of optimal classification line is solved. The main purpose of this classification interval is to maximize the classification accuracy of samples with two different distributions, and to minimize the error between the training data set and the actual value to zero, that is, to minimize the empirical risk. On this basis, the classification interval of statistical learning support vector machine (svm) model should be kept at the maximum, so that statistical learning support vector machine (svm) model algorithm has a better adaptive space when extrapolating the number of samples. The support vector machine is used to predict the time series. Starting from the definition of the initial time probability density function, the solution of the linear operator equation is obtained. In the process of classification and prediction, the statistical learning model of general support vector machine needs to extract information from global data and sample set features. In fact, the discriminant function only refers to the linear combination of internal product operations, that is, the internal product operations of unknown vectors and support vectors. Therefore, for the

classification problem, the computational complexity only depends on the number of support vectors.

When dealing with multiple classification problems, the solution of support vector machine algorithm is generally divided into two ideas: one is to treat multiple classification problems as a combination of multiple binary classification problems, and then use binary classification problems to solve. The second idea is to modify the objective function.

3. Design of Data Analysis System

3.1. Functional Requirements

The mass data analysis system designed in this paper can provide the following functions: real-time data collection, mass data storage, calculation and analysis, visualization of mass data, monitoring of Scribe clusters and Hadoop clusters, and visualization of these server memory and network bandwidth indicators.

3.2. System Design

We use real-time Scribe to collect the data on each log service. When Scribe collects the data, it writes it to Hadoop's HDFS, which provides data storage services. MapReduce is suitable for the analysis of such a large amount of data, and HDFS can provide the underlying file system support for MapReduce programming framework. The data analysis task is submitted to the Hadoop cluster by the client and calculated by the cluster. The calculation results are displayed visually in MySQL. When encountering a large amount of data, we use the Hadoop based HDFS HBase to provide database services. The web server uses the corresponding excuse to call. We use the PHP excuse to access HBase.

In this part, the data processing system we designed mainly consists of four parts, namely, data collection module, Hadoop module, HBase module and alarm module. The specific module structure is shown in Figure 1.

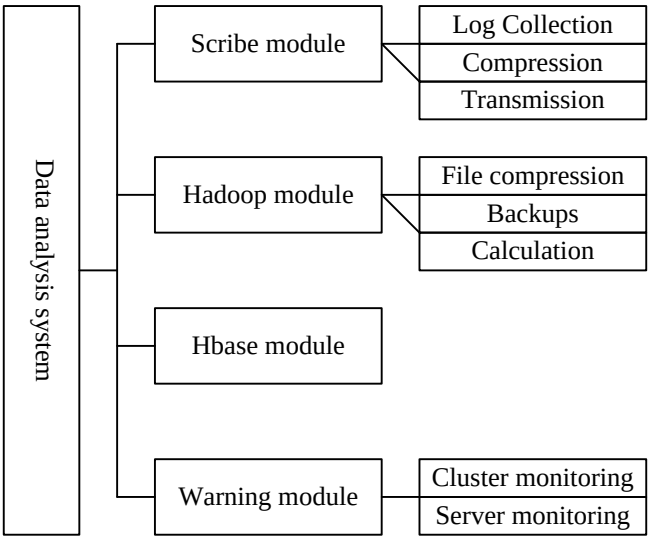

Figure 1. Basic composition of the data processing system

3.3. Testing of Data Analysis System

This paper mainly tests the accuracy of the original Scribe version and the compressed

multithreaded transmission version in collecting data of the same size and the same number of records, as well as the CPU and bandwidth usage. And test the following parameter settings of the compressed multi-threaded transmission plate. There are four test schemes, and the main test indicators are:

The theoretical value represents the relevant value of the original data file.

The actual value is the relevant value transferred to HDFS on Scribe.

CPU, mainly the percentage of CPU used.

The actual bandwidth represents the actual entrance bandwidth of the Scribe server.

Bandwidth, file size/transmission time required for file transmission.

4. Analysis of Test Results

4.1. CPU Test Results of Test Cases

In X1, when the client writes data to the Scribe local server, the CPU usage of the client program is high, while the Scribe local server only stores and forwards data, which is relatively low. The Scribe center server only receives and writes data to HDFS, and the CPU utilization is relatively low. The reason for the increased CPU utilization of the Scribe center server in test case X2 is the new compression function. See Table 1 for details:

Table 1. The CPU test results for the test cases

	Client	Scribe local server	Scribe center server
X1	72.67%	45.24%	34.11%
X2	72.89%	76.79%	58.97%
X3	81.29%	159.66	152.3%
X4	79.74%	95.37%	84.44%

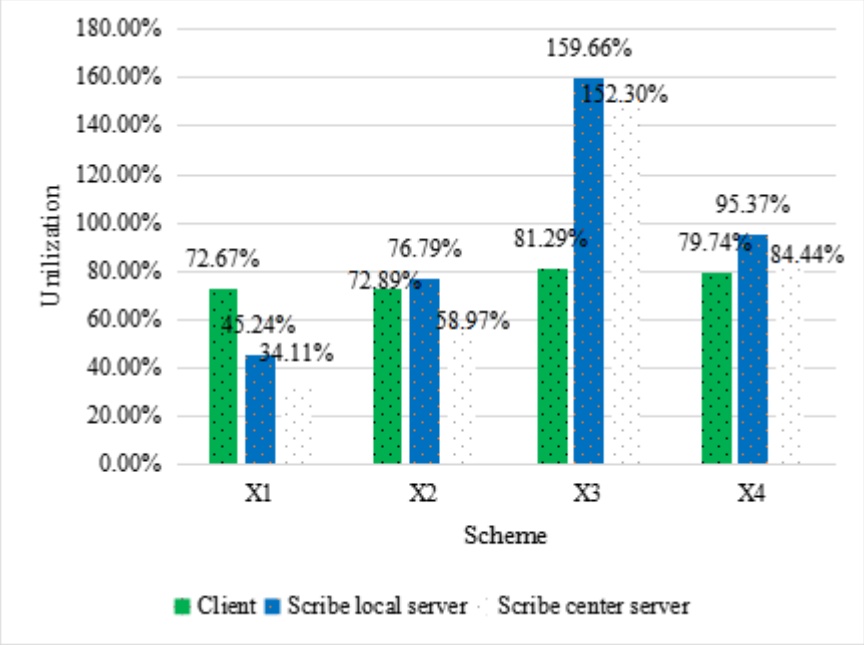

Figure 2. The CPU test results for the test cases

As shown in Figure 2, we can see that in X3, Scribe local server increases the CPU compression cost. The size of the packets received by the Scribe center server increases, resulting in an increase

in CPU overhead during decompression. In X4, the CPU utilization of the Scribe local server decreases because fewer records are compressed at the same time, which reduces the CPU used for compressing data.

4.2. Map Reduce Framework Efficiency Analysis

We analyze the computing efficiency of MapReduce framework based on the use of data with different data sizes as inputs. With the increase of data volume, the running time of Huamu tends to be linear. It is consistent with the growth of our input data. See Table 2 for details:

Table 2. MapReduce framework efficiency analysis

	Linearity	Runtime
5	100	50
10	200	200
30	600	680
35	700	680
50	900	880
60	1100	990

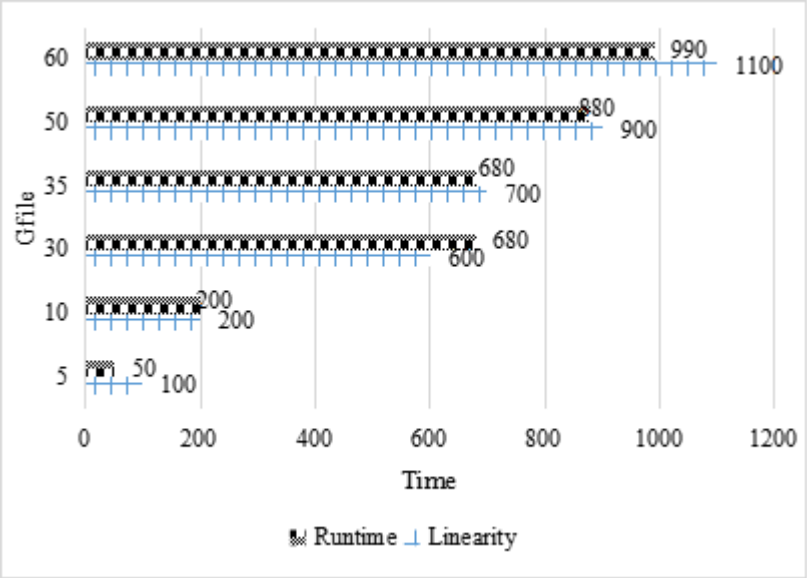

Figure 3. MapReduce framework efficiency analysis

As shown in Figure 3, we can see that when the data volume is 10 and 35, the time required for both operations is basically the same. This is because in our basic configuration, the file block size is 128M, and the input file can be divided into 8 blocks. When resources are fully utilized, the result is that the input sizes of the two are different, but the running time is basically the same. This also shows that it is more suitable for processing large files than smaller files.

5. Conclusion

In today's data mining technology, machine learning is one of the most important and basic parts. In data analysis, we can obtain useful information through a variety of methods. Therefore, effective algorithms are also needed to deal with these redundant and disorderly, low quality, incomplete expression or data problems that cannot be accurately calculated. In this paper, we

mainly study the precompression and denoising of original text data based on machine learning. The experiment shows that the size of the file also affects the ability of data processing.

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] K. Sailaja Kumar, H. K. Manoj, D. Evangelin Geetha:Twitter Data Analysis Using Hadoop and 'R' and Emotional Analysis Using Optimized SVNN. Comput. Syst. Sci. Eng. 44(1): 485-499 (2023).

[2] Matin N. Ashtiani, Bijan Raahemi: Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. IEEE Access 10: 72504-72525 (2022).

[3] Yu Kimura, Tatsunori Seki, Satoshi Miyata, Yusuke Arai, Toshiki Murata, Hiroyasu Inoue, Nobuyasu Ito:Hotspot analysis of COVID-19 infection using mobile-phone location data. Artif. Life Robotics 28(1): 43-49 (2023).

[4] Mohammad Izadikhah, Reza Farzipoor Saen:Developing a linear stochastic two-stage data envelopment analysis model for evaluating sustainability of supply chains: a case study in welding industry. Ann. Oper. Res. 322(1): 195-215 (2023).

[5] Dominik Raabe, Reinhard Nabben, Daniel Memmert:Graph representations for the analysis of multi-agent spatiotemporal sports data. Appl. Intell. 53(4): 3783-3803 (2023).

[6] Saad M. Darwish, Reham M. Essa, Mohamed A. Osman, Ahmed A. Ismail: Privacy Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics. IEEE Access 10: 76268-76280 (2022).

[7] Seyed Hossein Razavi Hajiagha, Hannan Amoozad Mahdiraji, Shide Sadat Hashemi, Jose Arturo Garza-Reyes, Rohit Joshi:Public Hospitals Performance Measurement through a Three-Staged Data Envelopment Analysis Approach: Evidence from an Emerging Economy. Cybern. Syst. 54(1): 1-26 (2023).

[8] Mohammad Mehdi Hosseinzadeh, Sergio Ortobelli Lozza, Farhad Hosseinzadeh Lotfi, Vittorio Moriggia:Portfolio optimization with asset preselection using data envelopment analysis. Central Eur. J. Oper. Res. 31(1): 287-310 (2023).

[9] Sana Khanam, Safdar Tanweer, Syed Sibtain Khalid:Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis. Comput. J. 66(1): 35-46 (2023).

[10] Irani Hazarika, Anjana Kakoti Mahanta: Mining Maximal Frequent Rectangles. Adv. Data Anal. Classif. 16(3): 593-616 (2022).

[11] Ali Hamdi, Khaled B. Shaban, Abdelkarim Erradi, Amr Mohamed, Shakila Khan Rumi, Flora D. Salim: Spatiotemporal Data Mining: a Survey on Challenges and Open Problems. Artif. Intell. Rev. 55(2): 1441-1488 (2022).

[12] Clarisse Dhaenens, Laetitia Jourdan: Metaheuristics for Data Mining: Survey and Opportunities for Big Data. Ann. Oper. Res. 314(1): 117-140 (2022).

[13] Tin C. Truong, Hai V. Duong, Bac Le, Philippe Fournier-Viger, Unil Yun: Frequent High Minimum Average Utility Sequence Mining with Constraints in Dynamic Databases Using Efficient Pruning Strategies. Appl. Intell. 52(6): 6106-6128 (2022).

[14] Tamil Selvi M, Jaison B: Lemuria: A Novel Future Crop Prediction Algorithm Using Data Mining. Comput. J. 65(3): 655-666 (2022). https://doi.org/10.1093/comjnl/bxaa093

[15] Durgesh Samariya, Jiangang Ma: A New Dimensionality-Unbiased Score for Efficient and Effective Outlying Aspect Mining. Data Sci. Eng. 7(2): 120-135 (2022).

[16] Esther Galbrun: The Minimum Description Length Principle for Pattern Mining: a Survey. Data Min. Knowl. Discov. 36(5): 1679-1727 (2022).

[17] Steedman Jenkins, Stefan Walzer-Goldfeld, Matteo Riondato: SPEck: Mining Statistically-Significant Sequential Patterns Efficiently with Exact sampling. Data Min. Knowl. Discov. 36(4): 1575-1599 (2022).

[18] Dmitrii Egurnov, Dmitry I. Ignatov: Triclusters of Close Values for the Analysis of 3D Data. Autom. Remote. Control. 83(6): 894-902 (2022).