Journal of Network Computing and Applications (2021) 6: 1-5
Clausius Scientific Press, Canada

DOI: 10.23977/jnca.2021.060101
EISSN 2371-9214

Research status and future prospects of machine learning algorithm in big data analysis

Haonan Wang

Data Science Experiment Center of College of Science, Liaoning University of Engineering and Technology, Fuxin 123000, China

Keywords: Big data, machine learning, naive Bayesian, k-means, SVM

Abstract: In the field of big data analysis, with the help of machine learning algorithm, the traditional mode of data analysis has been changed and the development of data processing and analysis has been promoted. This paper first explains the concepts of machine learning and big data analysis, then introduces the classic concepts and application concepts of three kinds of classical machine learning algorithms, namely naive Bayesian algorithm, K-means algorithm and SVM algorithm, and then discusses the application mode of machine learning algorithm in the emerging stage of big data environment, so as to enhance the value of big data processing. Finally, it looks forward to the big data processing The research trend of machine learning algorithm in this field is analyzed.

1. Forward

In the field of big data, researchers need to deal with data with large scale and complicated features. Because of the advantages of high efficiency of machine learning model and strong logic of algorithm model, machine learning has become the main way of intelligent analysis of big data at present.

2. The concept introduction of machine learning algorithm and big data analysis field

(1) Overview of machine learning algorithms

Machine learning is an important branch of computer science based on pattern recognition and artificial intelligence computational learning theory. The relevant research results show when the researchers deal with the larger the data scale, the efficiency of the machine learning model becomes higher [1]. Generally speaking, the central idea of machine learning is how to simulate and realize human learning behavior and how to explore the knowledge and skills acquired by computers. The accumulation of data is conducive to the improvement of classifier performance.

(2) Overview of the hierarchical process of big data processing

In emerging fields such as data mining, document classification indexing, the data objects faced by difficult problems are often huge and complex data sets. The big data processing flow as shown in Figure 1:

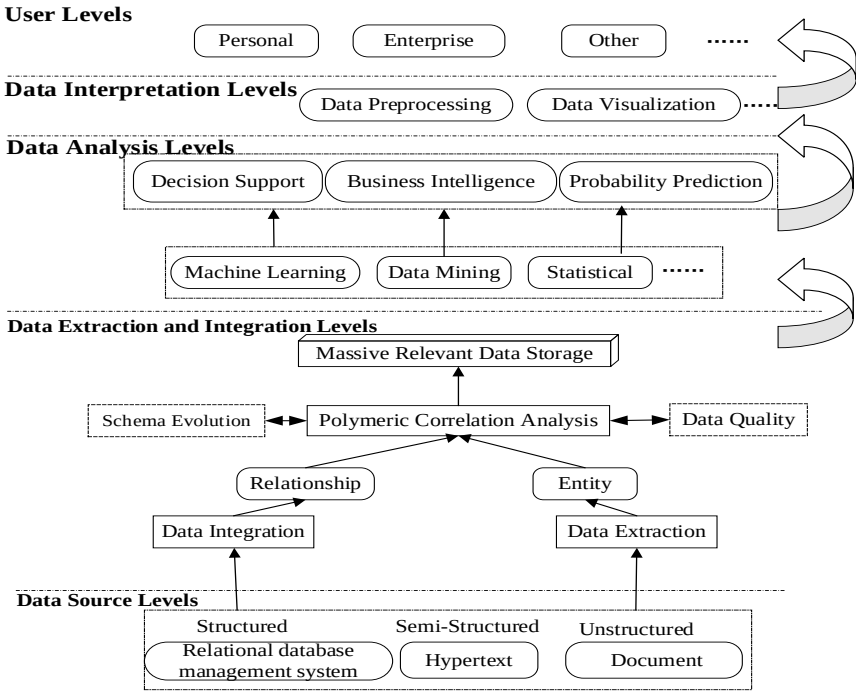

Figure 1: Big data processing hierarchy process

Among them, the core of big data processing flow is data analysis and data extraction and integration. The data analysis level uses non-traditional tools to deal with a large number of structured, semi-structured and unstructured data, so as to obtain a series of data processing results [2]. The data extraction and integration level deals with the data extracted and integrated from heterogeneous data sources.

(3) Overview of the definition of big data

The concept of big data analysis refers to the data collection that is perceived, acquired, managed, processed and served by software and hardware tools in a limited time, so as to obtain the information behind the data.

Data scientists in recent years have summarized a variety of different big data characteristics, and the specific meanings of the main characteristics are shown in Table 1:

Table 1: Big data feature map

Name	Name Meaning
Scale	Storage space range
Diversity	Types of various formats
Effectiveness	A timely limited time
Fuzziness	Methods of data collection

3. Introduction to main algorithms of machine learning

3.1 Naive Bayesian Algorithm

The principle of naive Bayesian algorithm is a supervised learning classification algorithm based on the assumption of independence of feature conditions. In the problem definition, a set of data sets

is given in the database. The naive Bayes algorithm can obtain the joint probability distribution under the environment of input and output ratio. Then, on the basis of statistical data, according to the conditional probability formula in practical inference, the conditional probability estimation of each feature attribute is obtained, and the probability that the current feature sample belongs to a certain classification is calculated, the specific calculation process is divided into four steps

1) Firstly, the feature attributes of a single object in the dataset are combined into a set of items to be classified

$$x = \{a_1, a_2, \cdots, a_m\} \tag{1}$$

2) Reorganizes a collection with categories into a new set

$$S = \{y_1, y_2, \cdots, y_n\} \tag{2}$$

3) The conditional probabilities are calculated separately

$$P(y_1 \mid x), P(y_2 \mid x), \cdots, P(y_n|x) \tag{3}$$

4) The final formula

$$P(y_i \mid x) = \frac{P(x \mid y_i)P(y_i)}{P(x)} (i = 1, 2, 3, \cdots, n) \tag{4}$$

In practice, there are some defects, such as the lack of training sample set, the interference of feature attributes and so on. The independent assumption of naive Bayesian algorithm can not be satisfied, so its performance is slightly worse than other machine classical learning algorithms, but more because of its simple implementation and low computational complexity, it has a very wide range of applications in text classification, network analysis and other fields [3].

3.2 k-means algorithm

In the field of machine learning, data similarity is an important research direction, that is, to find a data set with high similarity with the original data, and only accurately analyze the similarity between data [4]. K-means algorithm is a common clustering algorithm, which innovates a kind of indirect clustering method based on similarity measurement between different data sets. The specific algorithm model process is divided into three parts

1) Firstly, K value is randomly selected as the initial center by using some discrete objects in the data set.

2) According to the distance between some discrete objects and the cluster center, the part of objects are divided into the nearest cluster centers, and the model recalculates the clustering mean value to get a new cluster center

3) Repeat the second step until all the cluster centers are no longer changed, and the data set still tends to converge.

K-means algorithm cannot ensure the global optimal processing, but because of it is very simple, there is only one parameter to be adjusted in the model, which has excellent speed and good scalability. Therefore, as a classical clustering algorithm, which is widely used in the field of data science.

3.3 SVM algorithm (support vector machine)

SVM algorithm (support vector machine) is a machine learning algorithm based on statistical learning theory and the principle of structural risk minimization. It is a generalized linear classifier for binary classification of data, and its decision boundary is the maximum margin hyperplane

obtained by solving the learning samples. The basic idea is as follows: try our best to make the number set of the two categories have the maximum interval, find some data on the edge of the set, find a plane, that is, the decision surface of the practical problem, and finally get the maximum distance between the support vector and the decision surface.In order to make the new samples have classification and prediction ability, and make the data points closest to the separation surface have the maximum distance. The summary is to find support vectors iteratively on the basis of quadratic programming.The model structure is adjusted automatically by controlling parameters to minimize the empirical risk and structural risk. It is widely used in f many other fields.

4. Application mode of machine algorithm in the field of big data analysis

(1**)** Big data common processing mechanism mode

In the field of data mining, data objects are often large data sets. Traditional data processing methods are difficult to fully meet the requirements of big data processing. One of the important reasons is that the development of machine learning algorithm has not achieved parallel data processing.

It is with the core idea of parallel algorithm, combined with the core machine learning algorithm of different algorithm models dealing with the same problem, the large data sets in the database are processed in blocks, and the overall layout of big data is realized with the help of the results of each data.

(2**)** Divide and conquer strategy and sampling analysis mechanism

In the field of big data analysis, divide and conquer algorithm has always been an important computing paradigm. In the huge samples, the representative samples are selected according to the performance standards to form a subset, the distribution, topological structure and classification accuracy of the samples should be guaranteed to ensure the accuracy of the data of the subset samples. Finally, data analysis is carried out on this subset, namely big data divide and conquer strategy and sampling algorithm. With this processing method, the processing target of machine learning algorithm sample can be defined, and the machine can generate more accurate judgment. When selecting sample data, probability theory and compressed nearest neighbor method can be used to select the smallest data set corresponding to the large data set. Before the introduction of divide and conquer algorithm, it is still necessary to have the confidence level to meet certain requirements, carry out data elimination and data screening within the scope of credibility, and continuously optimize and improve the sample set.

5. Conclusion and future prospects

The appearance of AlphaGo, developed by Google, in the 2017 Go Game made more people pay attention to the field of machine learning and big data analysis. Although the classic algorithm of machine learning introduced above is relatively simple, it is the basic core of the development of machine learning. Data in the field of big data analysis is characterized by sparse feature attributes and complex data relationships. It needs machine learning algorithm to meet the needs of society. The combination of the two can not only improve the machine learning algorithm itself, but also make up for its own shortcomings. It can comprehensively improve the data processing ability.

The future prospect of machine learning in the field of big data lies in many fields, expecting more new algorithms to appear in the machine learning algorithm model. Since the development of human civilization, abundant empirical knowledge has been accumulated. How to benefit the existing human knowledge base and integrate it into the existing machine learning framework will become an important research point, in order to approach the ultimate goal of artificial intelligence.

References

[1] Ou huajie. Overview of machine learning algorithms under the background of big data [J]. China informatization, 2019(04): 50-51.
[2] He Qing, Li Ning, Luo Wenjuan, Shi Zhongzhi. Overview of machine learning algorithms under big data [J]. Pattern Recognition and Artificial Intelligence, 2014, 27(04): 327-336.
[3] Xu hongxue, sun Wanyou, du yingkui, Wang anqi. a summary of classical algorithms of machine learning and their applications [J]. Computer knowledge and technology, 2020, 16(33): 17-19.
[4] Ou Huajie. Overview of machine learning algorithms under the background of big data [J]. China Informatization, 2019(04): 50-51