Overview of Online Learning Algorithms for Large Data Analysis

Yonghong Zhang

Jiangxi Vocational College of Mechanical & Electrical Technology, Nanchang, 330013, China

Keywords: stream data; big data analysis; online learning

Abstract: We have entered the era of big data in modern and contemporary times, and after entering the era of big data, more and more fields have to deal with massive and high-speed data in real time. How to extract these large data streams and effectively transform them into practical information is very important for all walks of life. The traditional batch machine learning technology, which we are familiar with, has many restrictive factors in the application and analysis of large data in the contemporary era, while the current online learning technology adopts the streaming computing mode. Its main advantage is that the real-time analysis and calculation of data can be carried out directly in memory, so as to achieve the streaming mode. Data learning provides advantageous methods and tools, and it also introduces the background and motivation of big data analysis. It shows the latest and classical online learning algorithms and methods. Therefore, this latest online learning system is very likely and hopeful to solve all kinds of big data mining tasks. Challenges, the main technology and content include these aspects: first, linear model online learning; second, non-traditional online learning methods. All these methods give more detailed models or pseudo-codes as far as possible, so as to discuss a large number of machine learning for large data analysis. Key issues in use.

1. The Characteristics and Advantages of Streaming Data in All Aspects

In recent years, with the rapid development of new information technologies, such as computers, social networks and Internet technology, it is good to promote the ever-increasing growth of a large number of data types in human society, thus further promoting the human society to enter the era of big data [1-2]. And what we often call big data owners One of the most important characteristics of the first kind of large batch data is to store and analyze the data first, and then to centralize the analysis and calculation of some static data stored. The most typical large batch data calculation is Hadoop, which is a distributed file system. Hadoop Distributed Filesystem (HDFS) is mainly responsible for the storage of static data, and then data analysis and value discovery through MapReduce. This is a very important part. However, with the emergence of large data in all walks of life, such as images, video data, Internet data and so on. A series of large data is increasing, which is more presented to you in a dynamic stream data form [3]. This is what we call the second stream data. Unlike the traditional large batch data we see, stream data has more obvious characteristics and advantages, such as disorder, infinity, real-time, volatility, and so on. Extremely prominent features. In the calculation of large data streams, there are sometimes minor drawbacks, that is, when the order of arrival of some data streams cannot be determined, it will also be impossible to store all data streams. Therefore, when some data streams are no longer stored, we can directly calculate the data in real time with the arrival of flowing data. And the first point. Comparing with the large data batch computing we mentioned, the current large data flow mode is not optimistic either. The results of calculation research and practical experience are relatively less than that required. In summary, how to analyze and calculate large data flow efficiently, timely and rapidly through the data analysis of innovative methods has become the current big data analysis technology. New challenges in the field of technology.

2. Online Learning Method for Large Data Analysis

(1)Using the criterion function of perceptron: This method can prove that the mode weight of the

  DOI: 10.25236/mmmce.2019.013

algorithm will change after a limited number of amendments, that is, to a solution vector w. Thus, the algorithm of perceptron is a process of reward and punishment, and this is the earliest online learning algorithm in machine learning, which can be said to be very important. It can also solve the linear separability problem. So we can say that the emergence of perceptrons will promote the development of machine learning, but this method also has the situation that when the sample is linear and inseparable, the perceptron algorithm will not converge, so we need to use 2.1 section of the core perceptron algorithm, such as the second order perceptron (sec). Ondorder perceptron [23] and confidence-weighted algorithm are two very effective methods. The first second-order perceptron is actually an extension of the perceptron's directness. The common method of prediction is to update the augmented matrix or the fixed-valued unit matrix. The most representative of this process is that the augmented matrix and the misdivided data are discharged side by side, so that the weight vectors are updated and replaced according to the perceptron. The most obvious feature of the second confidence weighting algorithm is to maintain the different confidence levels of each feature. When the weights are updated, the weights corresponding to the features with lower confidence will be updated faster. On the contrary, the weights corresponding to the features with higher confidence will be updated more slowly. Therefore, this model also introduces second-order information.

(2)Online Active and Passive Algorithms: Since 2000, convex optimization model has become the main direction and purpose of machine learning research because it can proxy the global search for optimal solutions and is easy to implement. In online passive-active algorithm [7], this algorithm is a very typical one. Moreover, the core idea of online passive-active learning algorithm based on convex optimization model is based on a sample support vector machine, then the maximum spacing constraints between support vector machines can be transformed. When the plane spacing between new samples and formula (3) is less than 1, it can be generated. In a word, the principle of online passive-active algorithm updating is that when the error of new data is small or no error, the algorithm will be updated passively or not, and when the error of new data is large or there are errors, the algorithm will be active. Dynamic updating, i.e. projecting to the nearest neighbor of an existing classifier.

(3)Common online sparse solution learning algorithms: Sparse learning algorithm is based on the rise of compressed sensing technology, and L1 norm minimal regularization and other two more important factors can be further concerned. One of the most basic principles is LASSO, which can learn the model in Form (2). The parameters are chosen as the characteristics. Moreover, when the L1 paradigm is minimized, its optimal value can only be obtained on the boundary (i.e. only on the boundary of the coordinate axis), so that the sparse solution can be obtained. Finally, when we train in batch, we usually use the whole training. The sparse solution can be obtained by the method, and the online sparse solution learning algorithm can be used.

3. Summary and Prospects

In summary, in the modern and contemporary era of large data, online learning algorithm using streaming computing mode has great advantages, such as strong expansibility, high performance, flexibility and convenience, and small theoretical error range, etc. These obvious advantages, and it is in the field of dynamic data-oriented processing. It has been widely used, which makes it irreplaceable. This paper mainly shows us the latest and classical online learning methods, and gives the pseudo-codes and models of our common methods in detail as far as possible. At the same time, it also discusses the large-scale orientation. Some important and key problems in the application and research of large-scale machine learning in data analysis are comprehensively analyzed. We believe that the research methods of online learning algorithms in the era of big data include the following points: (1) Starting from the way of extracting research data streams; This paper mainly studies the organization methods and schemes of stream data, so as to improve the processing speed of stream data better; (2) reasonably and effectively put forward some novel online learning models to achieve a better online learning framework for deriving algorithms with high performance; (3) can propose a distributed stream-based online learning framework. There are

some advantageous schemes such as data mining, but these schemes must be based on data flow; (4) To develop some more automated tools for online learning, so as to greatly promote the application of online learning algorithms.

References

[1] Multitask Accelerated Online Learning Algorithms for Large Data Stream [J]. Li Zhijie, Li Yuanxiang, Wang Feng, Kuangli. Computer Research and Development, 2015 (11).

[2] Big data bioinformatics analysis promotes the discovery of tumor molecular targets [J]. Golden Yan, Yu Yingyan. Chinese Journal of Gastrointestinal Surgery, 2015 (01).

[3] Large Data Flow Computing: Key Technologies and System Examples [J]. Sun Dawei, Zhang Guangyan, Zheng Weimin. Journal of Software, 2014 (04).

[4] Big Data Management: Concepts, Technology and Challenges [J]. Meng Xiaofeng, Kindness. Computer Research and Development, 2013 (01).

[5] Strategies for anthropomics research based on a new generation of high-throughput technologies [J]. Yang Xu, Jiao Rui, Yang Lin, Wu Liping, Li Yingrui, Wang Jun. Inheritance, 2011 (08).

[6] A data stream extraction algorithm of circular sliding window tightly coupled with buffer [J]. Zhan Ying, Wu Chunming, Wang Baojun. Journal of Electronic Science, 2011 (04).

[7] Prediction of variable and constitutive splicing sites in the human genome [J]. Zhang Lirong, Luo Liaofu, Xing Yongqiang, Jinhongying. Advances in biochemistry and biophysics, 2008 (10).