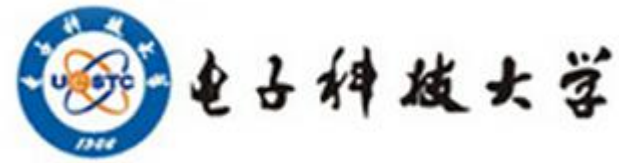

课程报告

Course Report

学　　院：信息与软件工程学院

学生姓名：杨庆

学　　号：201822090316

Machine Learning and Big Data

Abstract

Machine learning plays an increasingly important role in big data analysis. I mainly summarizes the methods and technologies of machine learning under the background of big data in this paper. Firstly, the basic model and classification of machine learning are briefly introduced. Then several key technologies of machine learning in big data environment are described. Then I shows several popular big data machine learning systems and analyzes their characteristics. Finally, I points out the main research directions of big data machine learning and the challenges it encounters.

1 Introduction

With the development of the network information, mobile internet, social network and e-commerce have greatly expanded the application field of the internet. We are in an era of "big data" with explosive data growth. Big data has a profound impact on people's lives. At the same time, the era of big data has brought new challenges and opportunities to human's ability to control data. We should use reasonable methods to analyze and process big data. Big data analysis mining processing is mainly divided into simple analysis and intelligent complex analysis. Simple analysis of commonly used SQL statements to complete some statistical and query work. However, the deep value of big data usually requires intelligent and complex analysis based on machine learning and data mining[1].

The core of big data is to utilize the value of data. Machine learning is the key technology to utilize the value of data. For big data, machine learning is indispensable. On the contrary, for machine learning, the more data, the more likely it is to improve the accuracy of the model. Therefore, the prosperity of machine learning cannot be separated from the help of big data. Big data and machine learning are mutually reinforcing and interdependent.

2. Research Status in Related Fields

Machine learning is a core research field of artificial intelligence. Machine

learning is a process of self-improvement by using the system itself. In this process, the performance of computer programs is continuously improved with the accumulation of experience. Experts and scholars have continuously proposed various learning task algorithms, which greatly improve the computer's ability to extract features from a large amount of data and discover hidden rules. Machine learning methods in data mining and analysis are increasingly widely used. Research shows that in many cases, the effect of machine learning model will be better with the larger the data processed.

In recent years, big data machine learning has become one of the research hotspots in the field of machine learning. Kleiner et al.[2] proposed a new data sampling method BLB (Bag of Little Boot-straps) based on the idea of Bagging in integrated learning, which is used to solve the bottleneck problem of Bootstrap when encountering big data. Shalev-Shwartz and Zhang et al.[3] proposed an improved gradient ascending (descending) method based on the idea of random learning to realize fast learning of large-scale models. Gonzalez et al. proposed a distributed machine learning framework GraphLab based on multi-machine clusters to realize large-scale machine learning based on graphs.

3. Overview of Machine Learning

Machine learning is a method that can endow machine learning with the ability to accomplish functions that cannot be accomplished by direct programming. However, from a practical point of view, machine learning is a method that uses data to train a model and then uses the model to predict.

First of all, we need to store historical data in the computer. Next, we process these data through machine learning algorithms. This process is called "training" in machine learning. The processed results can be used to predict new data. This result is generally called "model". The process of forecasting new data is called "forecasting" in machine learning. "Training" and "prediction" are two processes of machine learning, "model" is the intermediate output of the process, "training" produces "model" and "model" guides "prediction".

Human beings have accumulated a lot of history and experience in the process

of growth and life. Human beings regularly "sum up" these experiences and acquire the "laws" of life. When humans encounter unknown problems or need to "speculate" about the future, humans use these "laws" to "speculate" about unknown problems and the future, thus guiding their lives and work.

The "training" and "prediction" processes in machine learning can correspond to the "induction" and "speculation" processes of human beings. Through such correspondence, we can find that the idea of machine learning is not complicated, but only a simulation of human learning and growth in life. Since machine learning is not based on the results of programming, its processing is not causal logic, but a correlation conclusion drawn through inductive thinking.

This can also be associated with why human beings should learn history, which is actually a summary of human past experience. As a saying goes, "history is often different, but history is always strikingly similar." Through studying history, we can sum up the laws of life and country from history, thus guiding our next work, which is of great value. Some contemporary people ignore the original value of history and regard it as a means to publicize achievements, which is actually a misuse of the true value of history.

4. Classification of Machine Learning

Machine learning can be divided into supervised learning and unsupervised learning according to the learning form. Supervised learning is to give right and wrong instructions in the process of machine learning. Supervised learning is often used in prediction and classification. In supervised learning, a functional relation can be summarized from the trained data set, and then this functional relation can be used to predict new data and obtain results. In supervised learning, the training set needs to be input, then the target in the training set can be marked manually, and finally the output result can be obtained. Common supervised learning algorithms include statistical classification and regression analysis. Unsupervised learning, also known as inductive learning, is an algorithm to reduce errors and achieve classification through cyclic and decremental operations. Unsupervised learning has the highest intelligence but develops slowly, which is not the mainstream of current

research. In supervised learning, the unknown is often inferred from the known, which is risky and sometimes results are unreliable. Therefore, people have fully studied the first two and found a semi-supervised learning method, which has aroused great interest and concern.

5. Key Technologies of Machine Learning in Big Data

The most commonly used key technologies in machine learning include semi-supervised learning, ensemble learning, transfer learning, Bayesian network, decision tree, statistical learning theory and support vector machine, hidden Markov model, neural network, k-nearest neighbor method, sequence analysis, clustering, rough set theory, regression model, etc. In big data analysis, semi-supervised learning, transfer learning, probability graph model and integrated learning are especially important.

6. Research Direction and Challenges of Big Data Machine Learning

Big Data Machine Learning has two main research directions today. The first is the study of learning mechanism, which mainly studies how to make machines have some human behavior characteristics. The second is to study how to discover and mine valuable information in big data. People's research focuses mainly on the latter. Also, The field of machine learning will face the following challenges in the coming decades [4][5].

(1) Improvement of generalization ability of machine learning. This problem is very common. Generally speaking, multiple different objects have the same processing ability, which is called generalization ability. Currently, support vector machines have the strongest generalization ability.

(2) Speed problem. In the field of machine learning, people have been pursuing the goal of how to improve the speed of machine learning, the relationship between speed training and speed testing, and how to effectively weaken the conflict between the two are the issues that people are most concerned about.

(3) Understandability. In most cases, people can often get results through machine learning algorithms, but they don't know why they get such results. In the future big data analysis, more and more people hope to use machine learning

algorithm not only to get results, but also to know the reasons for the results.

(4) Data processing capability. In the past, most machine learning methods dealt with marked data. However, in the future big data analysis, machine learning algorithm will not only deal with a large number of unlabeled data, but also be interfered and influenced by some unbalanced data and garbage data.

7. Conclusion

There are many kinds of machine learning methods. Many machine learning methods need to be constantly revised and improved in order to play their role in big data analysis,. At present, big data machine learning system is still in an initial stage of exploration and research, and many aspects are not mature and perfect. In the era of big data, the analysis and mining of big data information cannot be separated from machine learning methods. With the development of big data, machine learning methods and machine learning systems will continue to develop. We can firmly believe that people will make full use of machine learning methods to obtain more and more useful information from big data in the future.

References

[1] Zhou, Z.H., Chawla, N.V., Jin, Y., et al. (2014) Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives. IEEE Computational Intelligence Magazine, 9, 62-74.

[2] Kleiner A., Talwalkar, A., Sarkar, P., et al. (2012) The Big Data Bootstrap. Proceedings of the 29th International Conference on Machine Learning (ICML), Edinburgh, 27 June-3 July 2012, 1759-1766.

[3] Bryant, R.E. (2011) Data-Intensive Scalable Computing for Scientific Applications. Computing in Science & Engineering, 13, 25-33.

[4] Darwiche, A. (2009) Modeling and Reasoning with Bayesian Networks. Cambridge University Press, Cambridge, 32-35.

[5] Pan, J.L. and Yang, Q. (2010) A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345-1359.