Big Data in Machine Learning

ZhiCheng Zhu
Indiana University Bloomington
936 S Clarizz Blvd
Bloomington, Indiana 47401
zhuzhic@iu.edu

ABSTRACT

With the development of IT technology, the world will enter the information age. People also called this "the era of third industrial revolution". Given the continuous development of the process of the third industrial revolution, all aspects of the traditional way of human society are changing. It can be said that every minute on the internet, have large amounts of new information be produced. With the increasing use of the internet and the increase of network bandwidth. People are making data every moment. It makes the information increasing quite quickly at a phenomenal rate. With so much data becoming available, getting data is not a problem for us anymore but find the right resource from the expanding information becoming a problem for most of the researchers. Because the data collected and stored at enormous speeds and human analysts may take weeks to discover useful information, traditional techniques are unfeasible for big data area. Therefore, we need to find new techniques to meet the challenge.

KEYWORDS

i523, hid229, Big data, Machine Learning, Technology

1 INTRODUCTION

The word "Machine Learning" first raised by Arthur Samuel, an American pioneer in the computer gaming and artificial intelligence areas [1]. In a broad sense, machine learning is a way to give the machine the ability to learn so that it can complete some task which not done directly by programming. It is a way of using Data to produce a model and then using the model to do prediction. Traditionally, if we want the computer to work, we will give it a series of computer instructions, and then the computer will follow this instruction step by step. The consequence always results by the code which you input before. And you can predict the result. But this way does not work in machine learning. Machine learning does not accept the instructions you entered at all, instead, it will accept the data you entered and applications of machine learning methods to these data sets and finally get produce a result. The core of big data is finding the value from the massive data sets, machine learning is a key technique that can effectively use the data value, for big data, the machine learning is indispensable. On the contrary, for machine learning, the more data will be more likely to improve the accuracy of the model. Therefore, the rise of machine learning is also inseparable from the help of big data. Big data and machine learning are mutually reinforcing and dependent.

2 CHALLENGES IN MACHINE LEARNING

Compared with the traditional machine learning, machine learning in big data can greatly expand the number of samples, the classification of many problems have a rich sample as support, this is the advantage of big data, but also caused many problems. Now, with the continuous optimization of hardware and programming algorithms, data collection and magnitude are no longer the major problems hindering the big data research. The relationship between different data sets, what kind of data is useful, which date is redundant and even cause interference to other data, how to reduce noise and make the model more accurate will be a challenge facing by machine learning when it mixes with big data. Big data has great potential value in all aspects of our society, it is not a simple task to obtain valuable information from big data. The core target of machine learning in big data is to excavate the data value which hidden in the data sets and find the information we need so as to maximize the value of data from the huge volume and structure of the data.

2.1 CHALLENGES IN PAST, NOW, AND FUTURE

Machine Learning has made extraordinary progress in the last 30 years. There have been significant advances in some area such as "Data Security, Finance, Marketing Personalization, Recommendations, and Smart Cars". But there are still a lot of obstacles for machine learning to go a step further. for example, lack of available data sets will be a problem whatever the past, the present, or the future. It has always been a problem. In the past, it is hard for a data scientist to find a tool to collect intensive data for researching. Nowadays, most of the data are collected by the big Internet company. the key problem of machine learning becomes to how people can get permission to access the data sets which owned by these big internet company. In the future, as the machine learning develop, the safety will become increasingly become the focus of attention. The Machine learning can be broken down into the following categories:

- "Supervised learning is a type of machine learning algorithm that uses a known data set (called the training data set) to make predictions. The training data set includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new data set. A test data set is often used to validate the model. Using larger training data sets often yield models with higher predictive power that can generalize well for new data sets." [3].

- "Unsupervised learning is a type of machine learning algorithm used to draw inferences from data sets consisting of input data without labeled responses" [4].
- Semi-Supervised Learning:A learning method combining supervised learning with unsupervised learning. Recognition is done using a large amount of unlabeled data and also using labeled data at the same time.

For example, one of the most common problem in Machine learning is Catastrophic forgetting, as we all know the machine learning need learning from the enormity of data and create a model. Catastrophic forgetting means the model will completely and abruptly forget previously learned information upon learning from some new data sets. If we want to achieve artificial general intelligence, then machine learning must be able to be used to perform multiple tasks. Even we can use representation learning and transfer learning to help us solve this problem to a certain extent but still has significant performance degradation. Another problem for machine learning is the safety. If we want to apply machine learning in peoplefis daily life. The security will be a question which the Data scientist unable to avoid. For example, in an image recognition test, "starting with an image of a panda, the attacker adds a small perturbation that has been calculated to make the image be recognized as a gibbon with high confidence"[6]. Another problem might raise because of the type of data sets. There are several different types of data in the Big Data area. The original unlabeled data and labeled data. the labeled will highly increase the efficiency of the process when the model starts to learning. Also, it will make the model more accurate when they do recognition. But the fact is even though the data increasing quite quickly at a phenomenal rate, with 2.5 quintillion bytes a day. but most of these data are unlabeled, which means these data are useless for supervised learning. and it also not suitable for deep learning which is the subset of machine learning.

3 APPLICATIONS IN MACHINE LEARNING

With the development expands and skills improved, machine learning becoming more popular and accepted by a lot of areas. Machine learning has been widely applied in data mining, computer vision, Natural Language Processing, biometrics, search engines, medical diagnosis and detection of credit card fraud, securities market analysis, DNA sequencing, speech and handwriting recognition, strategy games and robotics areas.

3.1 Machine learning in data mining

Data mining has been influenced by many disciplines, including database, machine learning, and statistics. To put it crudely, databases provide data management techniques, machine learning and statistics provide data analysis techniques.Many techniques provided by the statistical areas usually need further develop in the machine learning field, and then become effective machine learning algorithms before they can enter the field of data mining[5].Statistics affects data mining through machine learning, while machine learning and database are two major supporting technologies of data mining.

3.2 Machine learning in recommendation system

One of most common application in machine learning area is the recommendation system which running on the different Big internet company. Amazon may be taken as a typical example of the recommendation system. Based on a user's shopping record and a lengthy wish list, identifies which of the products the user is really interested in and willing to buy. Such a decision model can help the company to provide advice to customers and boost product consumption. for example, when you Log on to Facebook or GooglePlus, and they recommend the user who might be associated with you or you might know[2].

3.3 Machine learning in Marketing Personalization

According to the behavior pattern of the user during the free trial and the behavior in the past, which users might change to be a premium user, and which will not?. Such a decision model can help the company intervene in the program to convince users to pay sooner or better participate in product trials. For example, most of the video website and streaming media provider are willing to give user a free trial which can collect the user information and produce more attractive video or series to increase the user base[2].

4 CONCLUSION

Big data and machine learning are the two most popular fields in the Information Technology area. From the middle evil times' blocking of information to the explosion of data now, the amount of data in various fields and the scale of data sets have been increased at a phenomenal rate. The huge volume of data has brought huge potential opportunities and changes. With the proper use of the machine learning in big data can produce a lot of advantage. such as improve the efficiency. we can use the advantage of these data to help us make a better decision in different fields. one of a good example in scientific research is the data-driven research. In the scientific research, we can use the big data of the search engine to predict the ability widely used in the fields of medicine, astronomy and so on

ACKNOWLEDGMENTS

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper as well as TAs' helpful suggestions on this paper..

REFERENCES

[1] R. Kohavi and F. Provost. 1998. *Machine Learning*. Vol. 30. 271–274 pages.
[2] Bernard Marr. 2016. The Top 10 AI And Machine Learning Use Cases Everyone Should Know About. (2016). Retrieved October 09, 2017 from https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#7d4886ea94c9
[3] Mathworks. 2016. Introducing Machine Learning. (2016). Retrieved October 09, 2017 from https://www.mathworks.com/discovery/unsupervised-learning.html
[4] Mathworks. 2016. Introducing Machine Learning. (2016). Retrieved October 09, 2017 from https://www.mathworks.com/discovery/supervised-learning.html
[5] Margaret Rouse. 2017. data mining. (2017). Retrieved October 09, 2017 from http://searchsqlserver.techtarget.com/definition/data-mining
[6] Ophir Tanz and Cambron Carter. 2017. Why the future of deep learning depends on finding good data. (2017). Retrieved October 09, 2017 from https://techcrunch.com/2017/07/21/

why-the-future-of-deep-learning-depends-on-finding-good-data/

5 BIBTEX ISSUES

Warning–can't use both volume and number fields in Kohavi1

Warning–empty publisher in Kohavi1

Warning–empty address in Kohavi1

(There were 3 warnings)

6 ISSUES