Machine Learning under Big Data

Chunhe Shi[1, 2, a*], Chengdong Wu[3, b] Xiaowei Han[4, c]

Yinghong Xie[5, d] and Zhen Li[6, e]

[1, 3] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China

[2, 4, 5] College of Information and Engineering, Shenyang University, Shenyang 110044, China

[6]Northeast Regional Air Traffic Management Bureau of CAAC, Shenyang 110043, China

[a]schsydx@163.com, [b]wuchengdong@ise.neu.edu.cn, [c]hxw69@163.com,

[d]xieyinghong@163.com, [e]zlalbert@126.com

Keywords: Big data; Machine learning; Distributed and parallel computing; Classification.

Abstract. Currently big data is becoming the worldwide focus of attention, and using machine learning techniques to obtain valuable information from the massive data of complex structures has become a common concern yet an urgent problem. This paper analyzes and summarizes the present machine learning evaluation index under big data, and introduces some machine learning algorithms, then compares the differences between traditional algorithms and those under big data, and explores its developing trend.

Introduction

In recent years, as cloud computing, mobile Internet, and Internet of Things rapidly integrating, data increases exponentially. Hundreds of TB and even PB, EB-level data has excessively overloaded the processing capabilities of traditional computing techniques and information systems. How to learn and deal with these structured, semi-structured and unstructured large-scale massive data as well as the abilities of quickly acquiring valuable information becomes a problem worthy of attention yet of urgent solution. Machine learning is a core research area of artificial intelligence, whose theme is to imitate human learning activities. It studies methods of identifying current and acquiring new knowledge and improving qualities to realize self-perfection, Machine learning including Supervised Learning, Unsupervised Learning and Semi-Supervised Learning.

Development of data raises new demands and challenges towards machine learning from the aspects of research directions, evaluation indicators and key techniques. This paper mainly analyzes and summaries the present assessing indicators and algorithms, and then explores the evolving trend under big data.

The Machine Learning Evaluation Indicators under Big Data

Big data has low value density, thus complete sample is frequently adopted when data analyzing, which means large-scale data volume bring about unprecedented challenges to machine learning. Digging out value from data instead of being overwhelmed requires adaptability of machine learning techniques in the following aspects.

The Generalization Ability. The better generalization ability is usually expected for the machine learning algorithms trained through the samples, i.e. the capability of offering reasonable response to new input, which is the most important indicator to assess the performance of algorithms. The basic goal of machine learning is to generalize and popularize the training data instances.

The Comprehensibility. Many powerful algorithms are almost “black boxes”, whose users could only get the results but do not know why such results occur. With data volume growing and complexity increasing, users expect to get both the results and their producing procedure.

Capability of Data Utilization. Human ability of data collection is becoming stronger, and sorts of data collected are becoming various, including identified as well as massive unidentified data, and even inconsistent, incomplete dirty data containing noises. Discarding those dirty data as before and utilizing only those identified during the information process will definitely cause huge waste of data, meanwhile threatening the generalization ability of model acquired. Therefore, studying and developing machine learning methods that use all the data effectively is of great practical significance.

Machine Leaning Algorithms under Big Data

Due to the big volume and complexity of big data, traditional algorithms for small data are no more applicable in practice. Herein the study of algorithms under big data becomes a common concern for both academic and industrial fields. The machine learning under big data pays closer attention to the study of methods and algorithms. The Relationship between Machine Learning and Other Research Methods.

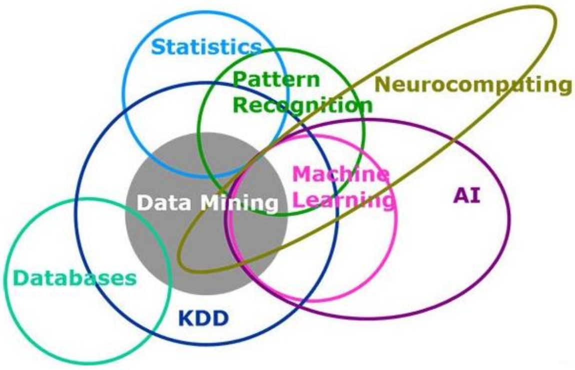

Figure 1. The Relationship between Machine Learning and Other Research Methods

The Decision Tree. Traditional decision tree algorithms, such as ID3, C4.5 and CART, etc. usually adopt the greedy top-down method. The core problem of decision tree is choosing the splitting property and its pruning. The best size of a tree is the adjustment parameters that can control the complexity of models, which should be determined by data adapting.

Traditional decision tree, as a classic classification algorithm, has a problem of large memory consumption when dealing with big data. Franco-Arcega [1] advanced a method of constructing decision tree from large-scale data to address some limitations in current algorithms, using all the data in training sets but not saving them in the memory. Experiments show that this method calculates faster in large-scale dealing. Ben-Haim [2] proposes an algorithm of constructing decision tree classifiers, which operates in a distributed setting and is suitable for large data sets and data streams. Compared with serial decision tree, this method can improve efficiency on the premise of approximate precision errors.

Artificial Neural Networks (ANN). Artificial neural network provides a popular and practical method and learn from the sample values for real, discrete or vector function. ANN learning has a good fitting effect for the training data. Many models have been put forward during the research of ANN, whose differences are mainly manifested in research approaches, network structure, operating mode, learning algorithms and their respective application. Neural network a learning algorithm based on empirical risk minimization, having some inherent defects, such as difficultly determinable layer and neuron number, being inclined to fall into local minimum and overfitting phenomenon, which could be well solved in SVM algorithms.

Recently, Huang etc. [3] has abandoned the iterative adjustment strategy of gradient descent algorithm, and put forward the Extreme Learning Machine (ELM), randomly evaluating the input weight parameters and bias of single hidden layer neural network (SHLNN), and working out by further calculation the output weight parameters. The training speed of ELM has been significantly increased.

SVM. Support Vector Machine (SVM) has a relatively better performance index [4], which is based on statistical learning theories. Via learning algorithm, the SVM can automatically pick out support vectors with better distinguishing capability of classifying, constructing out classifiers that maximize intervals between classes, and thus owning better adaptability and efficiency of distinguishing. SVM algorithm is aimed at finding a hyper plane H (d), which can separate the data in the training set, at the largest distance to the class field boundary perpendicularly. Thus SVM algorithm is also known as Maximum Margin algorithm. SVM algorithm ultimately eventuates to solving quadratic programming problem, and frequently used SVM algorithms comprise SVM-light, SMO, Chunking, etc.

Applying traditional statistical machine learning methods to big data classification involves two bottlenecks: a. computation-intensiveness, hardly applied to large-scale data sets; b. the unknown prediction about the fitting model of robustness and nonparametric confidence intervals. Towards these, Lau [5] provides an online learning algorithm, dealing with the classification of sequentially gradual input data, which provides faster computing speed, uses fewer support vectors and has better generalization ability. Pal [6], etc. put forward a feature selection method based on SVM for classification. As the accuracy of SVM classification algorithm concerns the feature number and the size of the data set, conducting feature selection before classification is beneficial to improve the accuracy.

Association Rule Algorithm. Analysis of association rules is a major type of machine learning tasks, expressing the relationship between two or more dichotomous variables by means of rules, and its successive development enabling the analysis of polychromous as well as continuous variables. Association rules compose no demand on data distribution, and the result is based on data without any subjective assumption, objectively reflecting the nature of data. Therefore, association rules have been widely applied to various fields after arising. Association rule algorithm is a solving process from input to output end.

There are two ways of solving the association analysis of big data: parallelization and increment. For parallelization, Li [7] put forward a parallel Apriority algorithm based on Map Reduce, whose main operation is to produce candidate item sets, parallelizing the process of producing the candidate item sets, improving the operating efficiency with better speedup ratio and scalability. Increment is mainly manifested in sequential pattern mining. [8] Advances an Incremental Sequence Mining (ISM) algorithm based on SPADE, which can not only maintain the frequent sequence during the database updating but also provide a user interactive interface to modify restrictions, e.g., Minus.

Machine learning algorithms also include Bayes Algorithm, EM Algorithm, Boosting and Baging Algorithm, etc, yet the specific introduction will not be presented here due to the limited space.

The Developing Trend of Machine Learning under Big Data

Ensemble Learning. Ensemble learning refers to integrating the results of different learning systems to obtain performance better than that of any single learning system. Even when adopting simpler learning system. Besides, the framework of ensemble learning essentially has the characteristic apt to parallelization, providing a good foundation for improving training and testing efficiency when dealing with big data. The principle of traditional machine learning is searching, searching the hypothetical spatial set composed of all the possible functions to find an approximate function closest to the unknown function. In general, the output result of traditional machine learning will face three problems: statistics, computation and representation.

In the era of big data, due to the large volume, complex structure and uneven quality of data, these problems may be more prominent; therefore, ensemble learning is bound to become a powerful tool of data analysis.

Transfer Learning. The ability of knowledge transferring and transforming between different scenarios is called transfer learning [9]. This is just what traditional machine learning lacks, which is rooted in traditional machine learning generally assuming that training and testing data obey the same data distribution, i.e. learned knowledge and applied problem have the same statistical characteristics. As the learning and applying scenarios transfer, statistical characteristics usually change correspondingly, greatly affecting the result of statistical learning.

The traditional machine learning algorithms usually just solve isolated tasks. Transfer learning attempts to transfer the knowledge learned in one or more source tasks to the related target tasks in order to improve the learning performance. Such techniques capable of transferring knowledge represent another step forward on the way from machine learning to human learning.

Parallelization and Distribution. As the parallel computation and the programming platform of big data occur and become more mature, Hardtop Map Reduce and Spark have currently been the mainstream platform of analyzing and processing big data. For the purpose of dealing with large-scale machine learning, much research work has been dedicated to accomplishing the design of various parallel machine learning and data mining algorithms based on Hadoop map Reduce, Spark and the traditional MPI parallel computing framework.

In order to directly use parallel machine learning algorithms, one way of the common practice is that in different parallel computing platforms professional designers realize parallel machine learning algorithms and provide a toolkit of machine learning and data mining, e.g. Mahout under Hadoop and Malibu under Spark. And yet for all the further study and exploration still remain to be devoted to parallelization and distribution.

Conclusion

In recent years, the development of big data pushes forward the progress of machine learning under big data and intelligence computing technology. Machine learning under big data is not merely a matter of machine learning and algorithm designing, but a matter of a large-scale and complex system. How to deeply analyze the complicated and varied data in order to efficiently use information is the present key direction of research. There are also many other research fields related to big data, and the further studying and exploration in the subsequent practical work is just awaiting.

Acknowledgements

This work was supported by the national natural science foundation of China under Grant（61503274)

References

[1] J.W. Han, K. Micheline. Data Mining: Concepts and Techniques, Vol. 2, Morgan Kaufmann Publisher (2006)

[2] X.D. Wu, X.Q. Zhu, G.Q. Wu, et al. Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering. Vol.26 (2014) No.1, p.97-107

[3] Mirjana M, Elisa Q, Letizia T. Data Mining for XML query-answering support. IEEE Transactions on Knowledge and Data Engineering, 2011.24(8):1393-1407

[4] Barwick H. The “four Vs” of Big Data. Implementing Information Infrastructure Symposium [EB/OL]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/

[5] Y.Q. He, Lee Rubao, Y. Huai et al. RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE).Hannover, Germany, 2011:1199-1208

[6] X L Dong, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. Proceedings of the VLDB Endowment. Vol.2 (2009) No.1, p.550-561.

[7] J.L. Liang, M.H. Zhang and X.Y. Zeng. Distributed Dictionary Learning for Sparse Representation in Sensor Networks. Image Processing, IEEE Transactions on Vol.23 (2014) No.6, p.2528-2541

[8] Schroeder WJ, Zarge JA. Lorensen WE. Decimation of triangle meshes. Computer Graphics, 1992.26(2):65-70.

[9] Manyika J. Chui M. Brown B. et al. Big data: The next frontier for innovation, competi tion, and productivity [EB/OL]. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[10] B. Huberman, “Sociology of Science: Big Data Deserve a Bigger Audience,” Nature, Vol. 482, (2012). p. 308.

[11] Ranjan, R. Modeling and Simulation in Performance Optimization of Big Data Processing Frameworks. Cloud Computing, IEEE. Vol.1 (2014). No.4, p.14-19

[12] Vaquero, L.M. Celorio, A. Cuadrado, F. et al. Deploying Large-Scale Datasets on-Demand in the Cloud: Treats and Tricks on Data Distribution. Cloud Computing, IEEE Transactions. 2014.3(2):132-144

[13] R. K. Lomotey and R. Deters. Analytics-as-a-service (AaaS) tool for unstructured data mining. In Cloud Engineering (IC2E), 2014 IEEE International Conference on, p.319–324. IEEE, 2014.

[14] K. Slavakis. G.B. Giannakis. G.Mateos. Modeling and Optimization for Big Data Analytics: (Statistical) learning tools for our era of data deluge. Signal Processing Magazine, IEEE. Vol.31 (2014) No.5, p.18-31

[15] H. Yu, T.J. Luo: NoSQL Database: A Scalable, Availability, High Performance Storage for Big Data. Pervasive Computing and the Networked World. 2014, p.172-183