IJCTA, 9(23), 2016, pp. 91-94
© International Science Press

Supervised Learning Techniques for Big Data: A Survey

G. Madhukar Rao* and Dharavath Ramesh**

Abstract : Information and Communication technology (ICT) produces the enormous data that lead to big data. The connection between the device to device communications generates a massive amount of data. There are many systems such as sensor networks, intrusion detection systems, climate change detection systems, satellite systems and health care systems are producing the massive amount of data, called big data. Significantly, it will lead to an exponential rise in computational complexity and unpredictability. This paper presents the new algorithmic approaches to handle the problems of big data, such as the minimal computational cost, reducing the complexity, overfitting and least memory requirements.

Keywords : ICT, Big data, complexity, Computational cost.

1. INTRODUCTION

Massive amounts of data are coming from the various sources that are very difficult to manage, process, and analyze through traditional database technologies [1]. In the instance of [2] and [3] defined big data as characterized by three Cs: volume, variety, and velocity. The term volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges.

Volume : It refers to the amount of data which are generated and stored from the various sources.

Variety : It refers to the various types of data collected via Information Communication Technology (ICT), Internet of Things (IoT), wireless networks and social networks.

Velocity : It refers to the speed of data generation, streaming and aggregation.

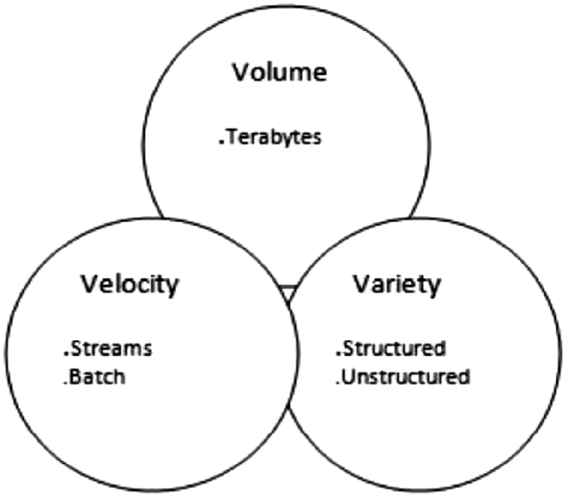

Fig. 1. Characteristics of Big Data.

* Department of Computer Science and Engineering IIT(ISM), Dhanbad, Jharkhand, India-826004 Email- madhusw511@gmail.com

** Department of Computer Science and Engineering IIT(ISM), Dhanbad, Jharkhand, India-826004 Email- ramesh.d.in@ieee.org

Big data analytics

An analysis of data can reveal many interesting properties of the data, which can help us to predict the future characteristics of data. Big Data analytics are the domain which explores the information, insights to the users. Machine learning techniques are used to classify and predict the future characteristics of data. This is a challenging task for many organizations, industries, institutions, health care sectors and government agencies. Extraction of useful information about the different types of data is the major problem in big data [4]. Big Data Classification is one of the research problem is used for extracting the information from the massive and complex data sets [5,6]. Data are coming from the different sources is called, Data Domain [12].

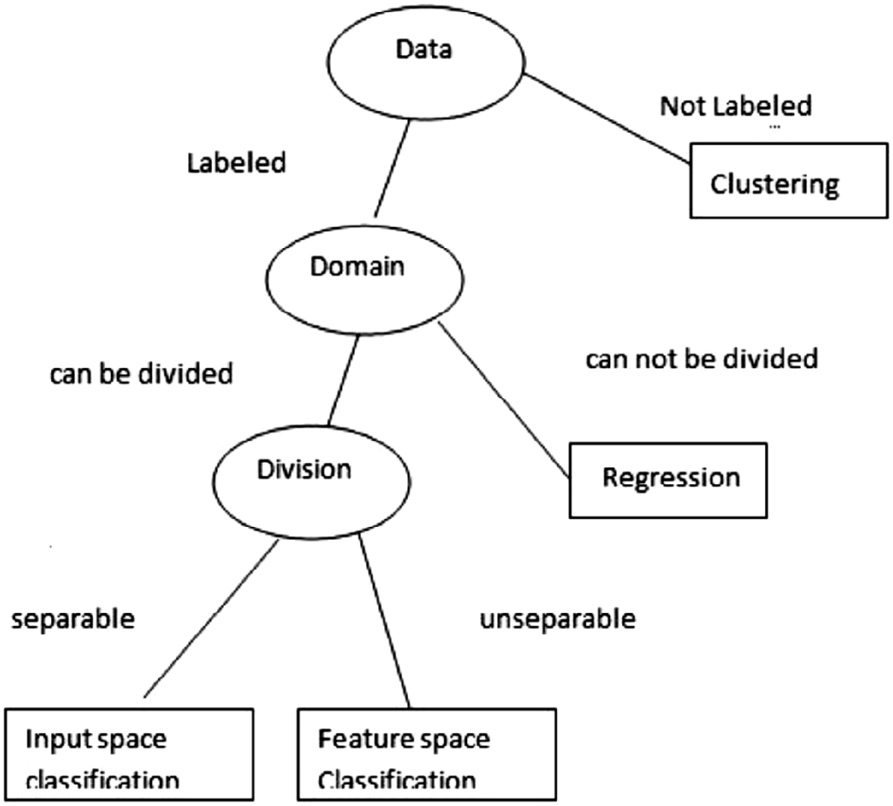

Fig. 2. Distinction between Regression, Classification and Clustering.

Domain perspective can help us to better understand of regression [7], classification [8], clustering [9], supervised learning [10], and unsupervised learning [11]. The figure 2 shows the distinction between regression, classification and clustering. Learning techniques are two types 1) Supervised Learning and Unsupervised Learning [26]. In supervised learning classes are known and classes and class boundaries are well defined in the given datasets and learning can be done from these classes, and it is called classification. In Unsupervised Learning classes are not known and class boundaries are not defined, the class labels themselves are learned and classes are defined based on that labels, and it is called it is called clustering.

2. CLASIFICATION TECHNIQUES

Classification models are classified into three types 1) Mathematical Model (*e.g.*, support vector machine) [13]. 2) Hierarchical Models (*e.g.*, Decision Tree and Random Forest) [14,15]. 3) Layered Models (*e.g.*, Deep Learning) [16].

Support Vector Machine

Traditional classification approaches are having poor performance when working with huge amount of data, but support vector machine is avoiding the problems of representing huge amount of data. Support vector machines are divided into linear and nonlinear SVMs [7]. It provides proper and accurate massive amount of data and compromising the classifier complexity and errors can be controlled explicitly.SVM[25] kernel can be applied directly to data without the need for the feature extraction process. SVM is an efficient model to handle the complex data and solving the problem of overfitting.

SVM[25] is not that much scalable, on large data sets because it takes a more amount of time for multiple scanning of data sets and it is too expensive to perform. Clustering based SVM (CB-SVM) is used to overcome the problems of SVM. It provides scalability and reliability of SVM classification [18]. CB-SVM is scalable when the efficiency of training, maximizing the performance of SVMs.

Decision Tree

Decision Tree can be used to filter the handle the massive amount of data. It can have satisfactory and accuracy of these data sets, which can be trained fast and provide sound results on those classifications of data [17].

The approach of Decision Tree is to break a large set of data set into n partitions, then learn a DT in each of the n partitions. Creation of final DT is pruning the tree removes nodes that do not provide accuracy in classification results due to reduced size tree. The limitations of Decision Tree are building a Decision tree is a time consuming process when the available data set is huge and the communication cost is minimized. C4.5 algorithm with map Reducing model can be used to overcome the limitations of Decision Tree [19].

Random Forest

Random forest is a supervised learning technique, which integrates sampling, subspace approach and an ensemble approach that consists of many decision trees. Random forest is fast to build decision trees and even fast to predict. Random forest is the technique, which is applied for both classification and regression problems [15, 7]. The decision Tree learning model has been injected into the Random Forest Models and thus includes the parameterization, optimization objectives of the decision tree learning model. The random Forest algorithm has balancing error in class population unbalance datasets. The advantage of Random forest is that it provides multiple trained decision classifiers for the testing phase. In which the generated forests can be saved for further use and it offers an experimental method for detecting variable interactions. Ovefitting is the one of the problems in Random forest for some datasets with noisy classification/regression tasks.

Deep Learning

One of the major challenges of Big Data is scalability of learning algorithms, especially scaling up of machine learning. Due to the rapid growth of the number of observations scaling-up problem may occur [20]. These kinds of problems can be handled by Deep learning [21], which is an alternative version of the artificial neural networks. Deep neural networks (DNNs) are having many hidden layers, whose weights are fully connected and often initialized Deep Belief Networks (DBMs) [22]. DNN has shown great performance in recognition and classification tasks, including natural language processing, image classification, and traffic flow detection [23]. Overfitting problem can be caused by incorrect values of the input parameters [7]. In deep learning models, the stochastic gradient descent(SGD) with a backward propagation approach has been used for learning parameters and scaling–up machine learning [24]. Overfitting problem can be exhibited from the integration of SGD based backward propagation technique. The advantage of Deep learning is to improve processing abilities, reducing the computational cost and complexity.

3. CONCLUSION

In this article, we have explored different types of supervised learning techniques for big data. All the techniques are suitable for different applications. Rapid Growth of data causes the high dimensionality, complexity, computational cost, overfitting and scalability problems. These problems can be handled by Machine learning algorithms for Big Data.

4. REFERENCES

1. Ibrahim Abaker Togo Hashem a, n, Ibrar Yaqoob a, Nor Badrul Anuar a, Salimah Mokhtar a, Abdullah Gani a Samee Ullah Khan b,The rise of “big data” on cloud computing: Review and open research issues,journal of information systems
2. P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, 2012.
3. J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp).
4. S. Bandari and S. Suthaharan. “Intruder detection in public space using suspicious behavior phenomena and wireless sensor networks,” in Proceedings of the 1st ACM International Work- shop on Sensor-Enhanced Safety and Security in Public Spaces at ACM MOBIHOC, pp. 3–8, 2012.
5. P. Zikopoulos, C. Eaton, et al. “Understanding big data: Analytics for enterprise class hadoop and streaming data.” McGraw-Hill Osborne Media, 2011.
6. S. Suthaharan. “Big data classification: Problems and challenges in network intrusion predic- tion with machine learning,” ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, pp. 70–73, 2014.
7. T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. New York: Springer, 2009.
8. S. Suthaharan. “Big data classification: Problems and challenges in network intrusion predic- tion with machine learning,” ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, pp. 70–73, 2014.
9. A.K. Jain. “Data clustering: 50 years beyond K-means.” Pattern recognition letters, vol. 31, no. 8, pp. 651–666, 2010.
10. S. B. Kotsiantis. “Supervised machine learning: A review of classification techniques,” Infor- matica 31, pp. 249–268, 2007.
11. [11]. O. Okun, and G. Valentini (Eds.), “Supervised and unsupervised ensemble methods and their applications,” Studies in Computational Intelligence series, vol. 126, 2008.
12. S.suthaharan,Machine Learning Models and Algorithms for big data classification, integrated series in information systems, DOI 10.1007/978-1-4899-7641-3_6.
13. M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. “Support vector machines.” Intelligent Systems and their Applications, IEEE, 13(4), pp. 18–28, 1998.
14. L. Rokach, and O. Maimon. “Top-down induction of decision tree classifiers-a survey.” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 35,vno. 4, pp. 476–487, 2005. 15. L. Bremen, “Random forests.” Machine learning 45, pp. 5–32, 2001.
15. L. Breiman, “Random forests.” Machine learning 45, pp.5–32, 200116. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R.
16. Salakhutdinov. “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580, 2012.
17. Dinxian wang .xiao liu, mengdi wang, “A DT –SVM Strategy for stock future prediction with big data ,”978-0-76955096-1/13,IEEE.
18. Hwanjo Yu,Jiong Yang, Jiawei Han,“classifying Large Data Sets Using SVMs with Hierarchical clusters,”SIGKDD’03 Washington, DC,USA, 1581137370/03/0008,2003,ACM
19. Wei dai,wei ji, “A MapReduce Implementation of C4.5 Decision Tree Algorithm,” International Journal of Database Theory and Application,SERSC,2014.
20. B. Dalessandro. “Bring the noise: Embracing randomness is the key to scaling up machine learning algorithms.” Big
21. Data vol. 1, no. 2, pp. 110–112, 2013.
22. L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. “Regularization of neural networks using drop connect.” In
23. Proceedings of the International Conference on Machine Learning, pp. 1058–1066, 2013.
24. L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inform. Process. 3 (2014), p. e2
25. Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with bigdata: a deep learning approach, IEEE Trans. Intell. Transp. Syst. PP (99) (2014)1–9..
26. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition.” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998
27. http://arxiv.org/ftp/arxiv/papers/1503/1503.07477.pdf.
28. http://link.springer.com/chapter/10.1007/978-1-4899-7641-3_1