2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

A SURVEY OF MACHINE LEARNING ALGORITHMS FOR BIG DATA ANALYTICS

Athmaja S.
Dept. of Computer Science
Bangalore University
Bengaluru, India

Hanumanthappa M.
Dept. of Computer Science
Bangalore University
Bengaluru, India

Vasantha Kavitha
Dept. of Computer Science
Maharani Lakshmi Ammanni College for Women
Bengaluru, India

Abstract*— Big data analytics is a booming research area in computer science and many other industries all over the world. It has gained great success in vast and varied application sectors. This includes social media, economy, finance, healthcare, agriculture, etc. Several intelligent machine learning techniques were designed and used to provide big data predictive analytics solutions. A literature survey of different machine learning techniques is provided in this paper. Also a study on commonly used machine learning algorithms for big data analytics is done and presented in this paper.

***Keywords*—big data, analytics, machine learning algorithms, technique, prediction, model**

I. INTRODUCTION

In this data rich era it is essential to use sophisticated analytics techniques on huge, diverse big data sets to produce useful knowledge and information. Big data analytics is a budding research area that deals with the collection, storage and analysis of immense data sets to trace the unknown patterns and other key information. Big data analytics helps us to recognize the data that are integral component to the future business decisions. Big data analytics can be abundantly found in domains such as banking and insurance sector, healthcare, education, social media and entertainment industry, bioinformatics applications, geospatial applications, agriculture etc. It is a herculean task to handle big data using conventional data processing applications. Thus to discover hidden data patterns, trends and associations, intelligent machine learning methods can be adapted. The objective of the current research paper is to discuss various machine learning algorithms used by data scientists for analyzing and modeling big data.

II. BIG DATA ANALYTICS

The term big data which describes extremely large data sets is widely being used among different researchers all over the world. Traditional relational databases are not capable of handling big data. Enormous quantity of data sets arrives from several sources like sensors, transactional applications, web and social media, etc. The big data phenomenon can be comprehended clearly by knowing the different V's associated with them- Volume, Velocity, Variety, Veracity and Value.

- Volume: This denotes the huge amount of data produced every second, oscillating between terabytes to zettabytes. These big data sets can be maintained using distributed systems.
- Velocity: This term represents the rate at which data is produced and processed to congregate the demands.
- Variety: This indicates the diverse range of data that we can use.
- Veracity: This speaks about the data quality. That is, it indicates the biases, noise, abnormality etc. in the data.
- Value: This points to the precious knowledge reveled from the data.

Data scientists use many well-marked analytics techniques. Text analytics, predictive analytics, natural language processing, machine learning, etc. are a few approaches to make better and faster decisions on big data sets to uncover hidden insights.

III. MACHINE LEARNING

Machine learning is an interdisciplinary research area which combines ideas from several branches of science namely, artificial intelligence, statistics, information theory, mathematics, etc. The prime focus of machine learning research is on the development of fast and efficient learning algorithms which can make predictions on data. When dealing with data analytics, machine learning is an approach used to create models for prediction. Machine learning tasks are mainly grouped into three categories- supervised, unsupervised and reinforcement learning. Supervised machine learning requires training with labeled data. Each labeled training data consists of input value and a desired target output value. The supervised learning algorithm analyzes the training data and makes an inferred function, which may be used for mapping new values. In unsupervised machine learning technique, hidden insights are drawn from unlabelled data sets, for example, cluster analysis. The third category, reinforcement learning allows a machine to learn its behavior from the feedback received through the interactions with an external environment [3]. From a data processing point of view, both supervised and unsupervised learning techniques are preferred for data analysis and reinforcement techniques are preferred for decision making problems [7].

978-1-5090-3294-5/17/$31.00 ©2017 IEEE

Most of the traditional machines learning algorithms were implemented for data sets which could be completely fit into the memory [15]. As the data keeps getting bigger day by day, many intelligent learning methods are being implemented to provide solutions to several big data predictive analytics problems. A study on several commonly used machines learning techniques for big data analytics is provided in the following section.

IV. LITERATURE SURVEY

J. Qiu et al. presented different machine learning algorithms for big data processing [7]. The first one is representation learning or feature learning which deals with learning data representations that make the data analysis process easier. It is found that the performances of the machine learning algorithms are strongly influenced by the selection of data representation (or features) [16]. This learning scheme plays a crucial role in dimensionality reduction tasks. The important steps under representation learning are feature selection, feature extraction and distance metric learning [14]. Feature selection (variable selection) techniques are used to find those features of data which are most relevant for use in model construction. Feature extraction techniques transform the high dimensional data into a low dimensional space. In distance metric learning, a distance function is constructed to calculate the distance between various points of a data set.

The authors mentioned about another hot learning technique called deep learning in their paper. Most of the ancient machine learning approaches follows shallow-structured learning architecture that containing a single layer of nonlinear feature transformations. Some of the examples of such learning techniques are Gaussian mixture models (GMMs), hidden Markov models (HMMs), support vector machines (SVMs), logistic regression, kernel regression etc. [9]. In contrast to the shallow- structured learning architecture, deep learning techniques make use of supervised and unsupervised strategies in deep architecture. The learning systems with deep learning architecture are composed of several levels of nonlinear processing stage, in which each lower layer's output is given as the input of the immediate higher layer. Some of the examples are deep neural networks, conventional neural networks, deep belief networks and recurrent neural networks etc. Because of the high performance of deep learning algorithms they are well suited for big data analytics applications.

Scalability is a challenging issue with the traditional machine learning algorithms. The traditional schemes cannot process the huge data chunks within a stipulated time as they require all the data in the same database. A new field of machine learning called distributed learning has been evolved to solve this problem. In this scheme, the learning is carried out on data sets distributed among several workstations to scale up the learning process [4]. Examples of the distributed machine learning algorithms are decision rules, stacked generalization, meta-learning and distributed boosting etc. Parallel machine learning is another popular learning scheme where the learning process is executed among multiple processor environments or on multiple threaded machines [1].

Transfer learning is another machine learning approach mentioned in their paper. A common practice is that both the training data and test data are taken from the same field in the conventional machine learning process. That is, the input feature space and data distribution are identical [8]. But there are certain scenarios in which getting training and test data from the same domain is a difficult and expensive task. In order to solve this issue, the transfer learning technique has been used. In this scheme a high performance learner is created for a target domain by getting trainings from a related source domain. Transfer learning techniques are widely being used in many real-world data processing applications.

The authors discussed about another learning scheme called active learning. In some cases the data is represented without labels which become a challenge. Manually labeling this large data collection is an expensive and strenuous task. Also, learning from unlabelled data is very difficult. Active learning is used to solve the above mentioned issue by selecting a subset of the most important instances for labeling [17]. Another scheme, kernel- based learning, has been widely used in many engineering applications to design efficient, powerful and high performance nonlinear algorithms [5]. Some of the algorithms capable of operating with kernels are support vector machines (SVM), principal component analysis (PCA), kernel perceptron, etc.

J.L. Berral-Garcia presented a paper describing the frequently used machine learning algorithms for big data analytics [6]. Several algorithms are used for performing modeling, prediction and clustering tasks. Decision tree algorithms (like CART, Recursive Partition Trees or M5), K-Nearest neighbors algorithms, Bayesian algorithms (using Byes theorem) , Support vector machines(SVM), Artificial Neural Network, K-means, DBSCAN algorithms, etc are presented in this paper. Several execution frameworks - Map-Reduce Frameworks (Apache Hadoop and Spark), Google's Tensor flow, Microsoft's Azure-ML were also mentioned. The implementations of the previously discussed algorithms are made available to the public through different tools, platforms and libraries such as R-cran, Python Sci-Kit, Weka, MOA, Elastic Search, Kibana etc.

M. U. Bokhari et al. presented a three layered architecture model for storing and analyzing big data [11]. The three layers are data gathering layer, data storing layer and data analysis & report generation layer. In order to gather and handle the huge volume of big data coming from high speed sources such as sensors or social media, a cluster of high speed nodes or severs are kept in the data gathering layer. The data storage layer is responsible for storing the big data. The Hadoop

Distributed File System (HDFS) can be used for data storage. In the data analysis layer, machine learning techniques such as ANN, Naive Bayes, SVM and Principal Component Analysis etc. are used to churn knowledge from the huge complex data chunks.

P.Y. Wu et al. in their paper provided case studies to show how big data analytics is useful in precision medicine to provide the most appropriate treatment to each patient [12]. Principal Component Analysis, Singular Value Decomposition and tensor-based approaches are useful for feature extraction and for feature selection filter based and wrapper based methods are helpful. All these are dimensionality reduction techniques. The authors compared different techniques for performing data mining tasks. Logistic regression, cox regression, local regression techniques are simple to interpret, but are prone to outliers. Logistic regression with LASSO regularization reduces feature space. But over fitting is a problem. Other models such as Hidden Markov models, Conditional Random fields, relational subgroup discovery, episode rule mining etc are also useful for performing data mining tasks. The authors discussed about the useful platforms for big data analytics. Apache Hadoop, IBM InfoSphere Platform, Apache Spark Streaming, Tableau, QlikView, TIBCO Spotfire, and other visual analytics tools are highly impactful platforms for providing big data analytics solutions. Two real world case studies such as integrative -omic data for the improved understanding of cancer mechanisms, and the incorporation of genomic knowledge into the EHR system for improved patient diagnosis and care were done to discuss the usefulness of biomedical big data analytics for precision medicine. Multi-omic TCGA [13] data and EHR data [2]were used to conduct this study.

M.R. Bendre et al. [10] conducted a research on the usage of big data in precision agriculture. The authors mentioned that big data provides a broad range of functions to uncover new insights to address several farming problems. The designed model uses the MapReduce technique for big data processing and the linear regression method for data prediction. The data collected from the KVR(Krishi Vidyapeeth Rahuri (KVR), Ahmednagar, India) station are used to test the model. The result forecasted using this model is very useful for effective decision making in the agriculture domain.

The following table summarizes the literature survey presented in this paper.

TABLE I. LITERATURE SURVEY SUMMARY

Sl. No.	Author(s) name (year)	Algorithms / Techniques	Summary
1.	J. L. Berral-Garcia (2016)	Decision tree algorithms, K- Nearest neighbor algorithms, Bayesian algorithms, SVM, ANN, K-means, DBSCAN	A survey was done on the various machine algorithms for classification, prediction and modeling
2.	J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng (2016)	Gaussian Mixture models, Hidden Markov Models, SVM, logistic regression, Kernel Rgression, Deep neural networks, Deep belief networks, PCA, Kernel Perceptron	A survey was done on the various traditional as well as advanced machine learning algorithms used for big data processing.
3.	M.U. Bokhari, M. Zeyauddin and M. A. Siddiqui (2016)	ANN, SVM,PCA, Naive Bayes	Presented a 3 layered architecture model for storing and analyzing Bigdata. Data storage can be done using the Hadoop Distributed File System(HDFS) and data analysis can be done using techniques like ANN, SVM, Naive Bayes and PCA.
4.	P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang (2016)	Logistic regression, PCA, HMM, Local regression, cox regression	Discussed several machine algorithms and platforms like Hadoop , IBM Infosphere, Tableu, Qlik view, Spark etc for providing big data solutions. Case studies were done using –omic data from TCGA and EHR data to show the usefulness of biomedical big data analytics for precision medicine.
5.	M. R. Bendre, R. C. Thool and V. R. Thool	MapReduce, Linear regression	A model was built using MapReduce and Linear regression techniques. Case study was carried out to predict the rainfall and temperature value for the year 2013 using the historical weather data collected from KVR, Ahmednagar. The main objective was to improvise the accuracy of rainfall forecasting.

V. CONCLUSION

With the advents in big data technology, it became difficult to handle the complex big data using the traditional learning algorithms. Therefore several advanced, efficient and intelligent learning algorithms are required to handle the huge chunks of heterogeneous datasets. The results obtained through these analytics techniques provide more effective solutions to many real world problems in various domains such as healthcare, agriculture, social media, banking etc. Various research papers are surveyed to gather information about advanced learning techniques. This paper gives an overall idea about the advanced machine learning algorithms

and techniques used to provide solutions to the big data analytics problems.

REFERENCES

[1] "Parallel machine learning toolbox", retrieved from http://www.research.ibm.com/haifa/projects/verification/ml_toolbox/.

[2] C. A. Caligtan and P. C. Dykes, "Electronic health records and personal health records", Semin Oncol Nurs, vol. 27, pp. 218-228, 2011.

[3] C.M. Bishop, Pattern recognition and machine learning, Springer, New York, 2006.

[4] D Peteiro-Barral and B Guijarro-Berdinas, "A survey of methods for distributed machine learning", Progress in Artificial Intelligence, Springer, vol. **2**, issue 1, pp. 1-11, 2013. DOI:10.1007/s13748-012-0035-5.

[5] G. Ding, Q. Wu, Y. D. Yao, J. Wang and Y. Chen, "Kernel-Based Learning for Statistical Signal Processing in Cognitive Radio Networks: Theoretical Foundations, Example Applications, and Future Directions", IEEE Signal Processing Magazine, vol. 30, issue. 4, pp. 126-136, 2013. DOI: 10.1109/MSP.2013.2251071.

[6] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", 18th International Conf. on Transparent Optical Networks, pp.1-4, 2016. DOI: 10.1109/ICTON.2016.7550517.

[7] J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing, Springer, vol. 2016:67, pp. 1-16, 2016. DOI: 10.1186/s13634-016-0355-x.

[8] K Weiss, T Khoshgoftaar and D Wang, "A survey of transfer learning*", Journal of Big Data, Springer, vol.* 3, issue 9, pp. 1-40, 2016. DOI: 10.1186/s40537-016-0043-6

[9] Li Deng, "A tutorial survey of architectures, algorithms and applications for deep learning", APSIPA transactions on Signal and Information Processing, vol. 3,pp.1-29,2014. DOI: https://doi.org/10.1017/atsip.2013.9.

[10] M. R. Bendre, R. C. Thool and V. R. Thool, "Big data in precision agriculture: Weather forecasting for future farming", 1st International Conf. on Next Generation Computing Technologies, pp. 744-750, 2015. DOI:10.1109/NGCT.2015.7375220.

[11] M. U. Bokhari, M. Zeyauddin and M. A. Siddiqui, "An effective model for big data analytics", 3rd International Conference on Computing for Sustainable Global Development, pp. 3980-3982, 2016.

[12] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang, "–Omic and Electronic Health Record Big Data Analytics for Precision Medicine", *IEEE Transactions on Biomedical Engineering*, vol. 64, issue 2, pp. 263-273, 2017. DOI: 10.1109/TBME.2016.2573285.

[13] T. C. G. Atlas. Available: http://cancergenome.nih.gov/.

[14] W. Tu and S. Sun, "Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives", Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining , ACM, pp. 18–25, 2012. DOI: 10.1145/2351333.2351336.

[15] X. W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives", in IEEE Access, vol. 2, pp. 514-525, 2014. DOI: 10.1109/ACCESS.2014.2325029.

[16] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, issue 8, pp. 1798-1828, 2013. DOI: 10.1109/TPAMI.2013.50.

[17] Y. Fu, B. Li, X. Zhu and C. Zhang, "Active Learning without Knowing Individual Instance Labels: A Pairwise Label Homogeneity Query Approach", in IEEE Transactions on Knowledge and Data Engineering, vol. 26, issue 4, pp. 808-822, 2014. DOI: 10.1109/TKDE.2013.165.