



MACHINE LEARNING CHALLENGES OF BIGDATA

N Sirisha^{1,2}, Sheikh Gouse¹,
K V D Kiran²

¹Department of Computer Science and Engineering,
MLR Institute of Technology, Hyderabad, India.

²Department of Computer Science and Engineering,
K L University, Vaddeswaram, Guntur District, India.

¹grandhishirisha@gmail.com,

²gouse.sheikh@gmail.com

June 11, 2018

Abstract

Machine learning sometimes refers to changes in systems that perform tasks related to Artificial Intelligence (AI). Such tasks involve recognition, diagnosis, planning, technology management, prediction, etc. also conferred. Machine learning is an intelligence methodology of discovering information for creating intelligent selections. Big Data has nice impacts on scientific discoveries and worth creation. This paper introduces ways in machine learning, main technologies in Big Data and a few applications of machine learning in Big Data. Challenges of machine learning applications in Big Data are mentioned. Some new ways and technology progress of machine learning in Big Data are also measured.

Keywords: Machine Learning, Big Data, Artificial Intelligence, Vs, Hadoop, HDFS, Map Reduce.

1 INTRODUCTION

Machine learning could be a branch in technology that studies the planning of algorithms that can learn. Typical machine learning tasks are unit thought learning, operate learning or predictive modeling, clump and finding prophetic patterns. These tasks are unit learned through on the market knowledge that was discovered through experiences or directions, as an example. Machine learning hopes that as well as the expertise into its tasks can eventually improve the learning. The final word goal is to enhance the educational in such the simplest way that it becomes automatic, in order that humans like ourselves do not have to be compelled to interfere to any extent further. Machine learning tasks are unit usually classified into 3 broad classes, counting on the nature of the educational "signal" or "feedback" on the market to a learning system. The basic machine learning techniques are:

1. Supervised learning[2]

It expects people to give both input and wanted output, not withstanding outfitting input about the exactness of forecasts amid preparing. When preparing is finished, the calculation will apply what was found out to new information. Supervised learning modifies or organizes parameters by an immediate examination between the genuine system yield and the coveted yield. It is a shut circle input framework, where the mistake is the criticism flag. The mistake measure, which demonstrates the contrast between the system yield and the yield from the preparation tests, is utilized to control the learning procedure.

2. Unsupervised learning

It should not be prepared with wanted outcome data. Rather, they utilize an iterative approach called deep learning to survey data and to reach conclusions. Its calculations are utilized for more hard assignments tasks than supervised. Its algorithms could contain imperfections in designs, with data focuses that are despicably associated and related, bringing about dishonorable order of similarities. The calculation, for instance, would not

appropriately arrange. No labels are unit given to the educational algorithmic rule, exploit it on its own to search out structure in its input. Unsupervised learning are often a goal in itself (discovering hidden patterns in data) or a way towards associate finish [3].

3. Reinforcement learning

It is a kind of Machine Learning and Artificial Intelligence. It enables machines and programming specialists to naturally decide the perfect conduct inside a particular context, so as to improve its execution. Example is Trojan horse and playing a game [2].

4. Semi-supervised learning

It is a combination of supervised and un-supervised, where the user provides associated trained and untrained data. Example is web page classification [3].

2 TECHNOLOGY MEASURES OF MACHINE LEARNING IN BIG DATA

Examples of supervised learning algorithms include Classification and regression algorithms, k-nearest neighbors, naivesbayes, discriminant analysis etc. Examples of unsupervised algorithms are clustering, association, k-means for clustering problems, Apriori algorithm. Examples of semi-supervised learning algorithms include all the supervised and unsupervised algorithms. Summary of several machine learning algorithms is shown in figure 1[10][11].

3 COMPARISON OF BIG DATA TECHNOLOGIES

The tools like Hadoop, MapReduce[34][35][36], Spark MLlib, FlinkMLlib, DryadLINQ, Apache Mahout, and IBM parallel machine learning [2].

Algorithms	type	Learning policy	Learning principles	Classification methods	characteristics
Naive Bayesian classifier	Generative	High likelihood estimation and posterior probability	Probabilistic computation	High posterior probability	Distributed and conditional independent assumption
Non-linear SVM	Discriminant	Loss of regular hinge (Min.). Soft margin (max.)	Sequential minimal optimization algorithm (SMO)	Maximum class of test samples	Super-plane separation, kernel trick
Linear SVM	Discriminant	Loss of regular hinge (Min.). Soft margin (max.)	Sequential dual method	weighted test sample (Max)	Super-plane separation
Stochastic gradient boosting	Discriminant	Addition minimization loss	Stochastic gradient descent algorithm	Linear combination of weights	Linear solution of weak classifier
Decision tree	Discriminant	Regularized maximum likelihood estimation	Feature selection, generation, prune	IF-THEN condition based on tree spitting	Classification tree

Figure 1: Summary of several machine learning algorithms

The machine learning techniques such as:

- Recommendation
- Classification
- Clustering

PivotalR is a machine learning tool on big data. It results the normal R user access to HDFS[35][36]. It also gives the access to the user to apply machine learning algorithms on big data [13]. There are so many technologies that supports Big Data analytics and applications. Figure 2 [12] compares different big data technologies. The table shows different types of systems and their comparative strengths and weaknesses.

The figure 3[15] [16] depicts two fundamental classifications of arrangements. The former classification gives the insights about the information, preparing, and calculation controls to deal with Big Data. The later classification portrays the creation and adjustment of various machine learning ideal models and the alteration of existing calculations for these paradigms[14].

	Main Memory Database	Multiple Parallel Processing database	Big Data Tool	Hadoop	NoSQL
Homogeneous	1	1	1	2	2
Accessibility	1	1	1	2	2
Faultlessness	1	1	2	1	1
Real-Time Transactions	1	1	1	3	3
Big Data Analytics	2	2	1	1	3
Big Data	3	2	2	1	1
Unstructured Data	3	3	2	1	1

Figure 2: Comparison of Big Data Technologies

Other machine learning classifications are Microsoft Azure[17], Google Cloud [18], Amazon [19] and IBM [20]. The programming tool are R dialect [21], MATLAB [22] and Weka [23].

Types	Approaches		Big Data Challenges									
			Volume	Variety	Velocity	Veracity	Validity	Vulnerability	Volatility	Visualization	Value	
Manipulation	Data	Dimensionality Reduction										
		Instance Selection	*	*		*	*	*	*	*	*	
		Data Cleaning										
	Processing	Vertical Scaling										
		Horizontal Scaling	*	*	*	*	*	*			*	
	Algorithm	Modifications										
		New Paradigm	*	*	*		*	*				
Learning	Deep		*	*	*		*		*		*	
	Online		*	*	*	*	*	*	*	*	*	
	Local		*	*	*						*	
	Transfer		*	*	*	*		*		*	*	
	Life Long		*	*	*	*		*			*	
	Ensemble		*		*					*	*	

- represent high degree of remedy
- * represents partial resolution

Figure 3: Machine learning approaches and the challenges they address

4 CONCLUSIONS

This paper gives a precise audit of the difficulties identified with machine learning with Big Data and sorted them as indicated by the ten 'V' measurements of Big Data. Also, it gives a diagram of

Machine Learning approaches and talked about how these systems conquer the different difficulties distinguished. Moreover, the relations amongst methodologies and difficulties give a more careful comprehension of Machine L with Big Data. In this way, it satisfies the main goal to make an establishment for a more profound comprehension of machine learning with Big Data.

References

- [1] Sirisha N, K.V.D Kiran,;PROTECTION OF ENCROACHMENT ON BIG-DATA ASPECTS. International Journal of Mechanical Engineering and Technology (IJMET) ,pp. 550558, Volume 8, Issue 7, July (2017).
- [2] Lidong Wang, Guanghui Wang, Cheryl Ann Alexander. “Machine Learning in Big Data”, International Journal of Advances in Applied Sciences (IJAAS).Vol. 4, No. 4, December 2015, pp. 117 123, ISSN: 2252-8814.
- [3] C. L. P. Chen, C. Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on BigData”, Information Sciences, Vol. 275, No. 10, pp. 314-347, August 2014.
- [4] K. M. Tarwani, S. S. Saudagar, H. D. Misalkar, “Machine Learning in Big Data Analytics: An Overview”,International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, No. 4, pp.270-274, April 2015.
- [5] U. Jaswant and P.N. Kumar, “Big Data Analytics: A Supervised Approach for Sentiment Classification UsingMahout: An Illustration”, International Journal of Applied Engineering Research, Vol. 10, No. 5, pp. 13447-13457,2015.
- [6] Y. Demchenko, P. Grosso, C. De Laat, P. Membrey, “Addressing Big Data Issues in Scientific DataInfrastructure”, 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA,USA, pp. 48-55, May 2013.

- [7] D. E. O'Leary, "'Big Data', the 'Internet of Things' and the 'Internet of Signs'", *Intelligent Systems in Accounting, Finance and Management*, Vol. 20, pp. 53-65, 2013.
- [8] H. V. Jagadish, A. Labrinidis, Y. Papakonstantinou, et al., "Big Data and Its Technical Challenges", *Communications of the ACM*, Vol. 57, No. 7, pp. 86-94, 2014.
- [9] A. Zaslavsky, C. Perera and D. Georgakopoulos, "Sensing as a Service and Big Data", *International Conference on Advances in Cloud Computing (ACC)*, Bangalore, India, pp. 1-8, July 2012.
- [10] M. Turk, "A chart of the big data ecosystem", take 2, 2012.
- [11] L. Li, "Experimental Comparisons of Multi-class Classifiers", *Informatica*, Vol. 39, pp. 71-85, 2015.
- [12] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, Vol. 31, pp.249-268, 2007.
- [13] J. Dean, "Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners", John Wiley Sons, Inc., 2014.
- [14] X.-W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives", *IEEE Access*, vol. 2, pp. 514525, 2014.
- [15] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep Learning Applications and Challenges in Big Data Analytics", *J. Big Data*, vol. 2, no. 1, p. 1, Feb. 2015.
- [16] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics", *J. Big Data*, vol. 2, no. 1, pp. 120, 2015.
- [17] P. D. C. de Almeida and J. Bernardino, "Big data open source platforms", in *Proc. IEEE Int. Congr. Big Data*, Jun. 2015, pp. 268275.

- [18] R. Barga, V. Fontama, and W. H. Tok, “Cortana analytics”, in Predictive Analytics With Microsoft Azure Machine Learning. Berkeley, CA, USA: Apress, 2015, pp. 279283.
- [19] Google. (2016). Google Cloud Machine Learning.[Online].Available:<https://cloud.google.com/products/machine-learning/>
- [20] Amazon Web Services. (2016). Amazon Machine Learning. [Online]. Available: <https://aws.amazon.com/machine-learning/>
- [21] IBM. (2014). IBM Watson Ecosystem Program.[Online].Available:<http://www03.ibm.com/innovation/us/watson/>
- [22] R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2015, vol. 1.
- [23] MATLAB, TheMathWorks Inc., Natick, MA, USA, 2016.
- [24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update”, ACMSIGKDD Explorations Newslett., vol. 11, no. 1, pp. 1018, 2009.
- [25] L. Deng and X. Li, “Machine learning paradigms for speech recognition: An overview”, IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 5, pp. 10601089, May 2013.
- [26] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 17981828, Aug. 2013.
- [27] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, vol. 13. Cambridge, U.K.: Cambridge Univ. Press, 2014.
- [28] L. Bottou and V. Vapnik, “Local Learning Algorithms”, Neural Comput., vol. 4, no. 6, pp. 888900, 1992.

- [29] K. Huang, H. Yang, I. King, and M. R. Lyu, "Local learning vs. global learning: An introduction to maxi-min margin machine", in Support Vector Machines: Theory and Applications. Berlin, Germany: Springer, 2005, pp. 113131.
- [30] L. Torrey and J. Shavlik, Handbook of Research on Machine Learning Applications and Trends. Hershey, PA: IGI Global, 2010.
- [31] D. L. Silver, Q. Yang, and L. Li, "Lifelong machine learning systems: Beyond learning algorithms", in Proc. AAAI Spring Symp., 2013, pp. 4955.
- [32] T. Dietterich, "Ensemble methods in machine learning", in Multiple Classifier Systems, vol. 1857. London, U.K.: Springer-Verlag, 2000, pp. 115.
- [33] M. Sewell, "Ensemble learning", Dept. Comput. Sci., UCL, London, U.K., Tech. Rep. RN/11/02, 2011, p. 12.
- [34] RaswithaBandi, Sheikh Gouse, J.Amudhvel. (2017) A Comparative Analysis For Big Data Challenges And Big Data Issues Using Information Security Encryption Techniques1, 2 in International Journal of Pure and Applied Mathematics, Vol.8 Pages183-189 ISSN 978-3-319.
- [35] Sheikh Gouse, RaswithaBandi, P.Amarendar Reddy. (2017) "Comprehensive Survey on Big Data Analytics and Tools", presented in the National Seminar on International Conference on Recent Trends in Computer Science and Technology (ICRTCST-2017) R.V.S College of Engineering and Technology Edalbera, Bhilai Pahari, Jamshedpur-831012, 21/04/2017.
- [36] Sheikh Gouse, RaswithaBandi, Antiha. (2017) "Superintendence of Big Data values and Challenges Shielded from Intrusion" presented in the National Seminar on First International Conference on Recent Innovations in Engineering and Technology (ICRIEAT-2016) is being conducted by Auroras Scientific, Technological and Research Academy, Hyderabad on 22 & 23 December, 2016.

