A Survey on Big Data Analytics: Methods, Opportunities and Challenges

Ms. Kiran R. Borade[1], Mrs. Dhanashree S. Joshi[2], Mrs. Tejal S. Sonawane[3]

Lecturer, Department of Computer Engineering[1,2,3]

Guru Gobind Singh Polytechnic, Nashik, Maharashtra, India

kiranborade@ggsf.edu.in[1], dhanashree.joshi@ggsf.edu.in[2], tejal.sonawane@ggsf.edu.in[3]

Abstract: *In this paper, we briefly introduce the basic concepts and features of big data. We are surrounded by a huge amount of data but in demand of information. In the Big Data era, how to quickly get high quality and valuable information from large amounts of data has become an important research tool. Big Data Analytics has become very popular, not only academically, but also in industry and government programs, which can be attributed to the fact that Big Data Analytics offers great promises and poses significant global challenges. The extracted data can be useful to the organization in various aspects. Many decisions have to be made by business organizations based on this big data. Data is growing at a tremendous rate these days and it is an important issue for data management and management to analyze the information needed to save time and cost. The main purpose of this paper is to present a detailed analysis of the various forums/techniques suitable for Big Data processing. In this paper, the various software technologies, methods and tools available for Big Data analysis are explored and analyzed in detail with their strengths and weaknesses. In this paper various data mining and feature extraction methodologies are studied and finally comparison between technologies or tools is done based on performance of various systems.*

Keywords: Big data technology, Data Mining, Big Data, Big Data Analytics, Feature extraction

REFERENCES

[1]. J. Abello, P. M. Pardalos, and M. G. Resende, Handbook of Massive Data Sets (Massive Computing). vol. 4, New York: Springer, 2002.

[2]. C. C. Aggarwal and P. S. Yu, Eds., Privacy-Preserving Data Mining: Models and Algorithms. New York: Springer, 2008.

[3]. C. C. Aggarwal, Data Streams: Models and Algorithms. vol. 31, New York: Springer, 2007.

[4]. (2013). Australian Dept. Immigration. Fact sheet 70— Managing the border Internet. [Online]. Available: http://www.immi.gov.au/media/fact-sheets/70border.htm

[5]. J. Bader and E. Zitzler, ―HypE: An algorithm for fast hypervolume-based manyobjective optimization,‖ Evol. Comput., vol. 19, no. 1, pp. 45–76, 2011.

[6]. P. Baldi and P. J. Sadowski, ―Understanding dropout,‖ in Advances in Neural Information Processing Systems 26. Cambridge, MA: MIT Press, 2013, pp. 2814–2822.

[7]. M. Banko and E. Brill, ―Scaling to very very large corpora for natural language disambiguation,‖ in Proc. 39th Annu. Meeting Association Computational Linguistics, Toulouse, France, 2001, pp. 26–33.

[8]. A.-L. Barabási, Linked: The New Science of Networks. New York: Basic Books, 2002.

[9]. M. Basu and T. K. Ho, Data Complexity in Pattern Recognition. London, U.K.: Springer, 2006.

[10]. Y. Bengio, ―Learning deep architectures for AI,‖ Foundations Trends Mach. Learn., vol. 2, no. 1, pp. 1–127, 2009.

[11]. E. Brynjolfsson, L. Hitt, and H. Kim. (2011). Strength in numbers: How does data-driven decision making affect firm performance? [Online]. Available: http://ssrn.com/abstract=1819486

[12]. T. Chai, Y. Jin, and S. Bernhard, ―Evolutionary complex engineering optimization: Opportunities and challenges,‖ IEEE Comput. Intell. Mag., vol. 8, no. 3, pp. 12–15, 2013.

[13]. N. V. Chawla, ―Data mining for imbalanced datasets: An overview,‖ in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. New York: Springer, 2005, pp. 853–867.

[14]. N. V. Chawla, Learning on Extremes-Size and Imbalance-of Data. Tampa, FL: Univ. South Florida, 2002.

[15]. Q. Da, Y. Yu, and Z.-H. Zhou, ―Learning with augmented class by exploiting unlabeled data,‖ in Proc. 28th AAAI Conf. Artificial Intelligence, Quebec City, QC, Canada, 2014.

[16]. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, ―A fast and elitist multiobjective genetic algorithm: NSGAII,‖ IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002.

[17]. T. G. Dietterich, ―Approximate statistical tests for comparing supervised classification learning algorithms,‖ Neural Comput., vol. 10, no. 7, pp. 1895–1923, 1998.

[18]. D. Easley and J. Kleinberg, Networks, Crowds, and Markets. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[19]. Farhangfar, L. Kurgan, and J. Dy, ―Impact of imputation of missing values on classification error for discrete data,‖ Pattern Recognit., vol. 41, no. 12, pp. 3692– 3705, 2008.

[20]. L. Feng, Y.-S. Ong, I. Tsang, and A.-H. Tan, ―An evolutionary search paradigm that learns with past experiences,‖ in Proc. IEEE Congr. Evolutionary Computation, Brisbane QLD, Australia, 2012, pp. 1–8.

[21]. M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, ―Mining data streams: A review,‖ ACM SIGMOD Record, vol. 34, no. 2, pp. 18–26, 2005.

[22]. W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, ―Onepass AUC optimization,‖ in Proc. 30th Int. Conf. Machine Learning, Atlanta, GA, 2013, pp. 906–914.

[23]. W. Gao and Z.-H. Zhou, ―Dropout rademacher complexity of deep neural networks,‖ CORR abs/1402.3811, 2014.

[24]. J. A. Hanley and B. J. McNeil, ―A method of comparing the areas under receiver operating characteristic curves derived from the same cases,‖ Radiology, vol. 148, no. 3, pp. 839–843, 1983.

[25]. G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, ―The ―wake-sleep‖ algorithm for unsupervised neural networks,‖ Science, vol. 268, no. 5214, pp. 1158–1161, 1995.

[26]. G. E. Hinton and R. R. Salakhutdinov, ―Reducing the dimensionality of data with neural networks,‖ Science, vol. 313, no. 5786, pp. 504–507, 2006.

[27]. H. Ishibuchi, M. Yamane, N. Akedo, and Y. Nojima, ―Many-objective and many-variable test problems for visual examination of multiobjective search,‖ in Proc. IEEE Congr. Evolutionary Computation, Cancun, QROO, 2013, pp. 1491–1498.

[28]. H. Ishibuchi and T. Murata, ―Multi-objective genetic local search algorithm,‖ in Proc. IEEE Int. Conf. Evolutionary Computation, Nagoya, Japan, 1996, pp. 119–124.

[29]. Y. Jin and J. Branke, ―Evolutionary optimization in uncertain environments—A survey,‖ IEEE Trans. Evol. Comput., vol. 9, no. 3, pp. 303–317, 2005.