Big Data: Data Science Applications and Present Scenario

Shubhankar Chaturvedi and Shwetank Kanava
Arya College of Engineering and Information Technology,Jaipur

Abstract

In this paper we are presenting some simple study of data science which has been discussed very frequently in scientific community. We are also giving some recent trends and techniques and their impact on scientific as well as social community.

I. INTRODUCTION

Data science is not a new concept or term for statistical scientists, but this is a simply computerization of statistical old methods as per the need of present time. Big Data are large data sets. However, there is a lot of informative value hidden in this data, so many companies have desire to access this data for maximize their profit. It is very innovative and competitive research area in present time.

II. DATA SCIENCE & CHALLENGES

During the last few years, the most challenging problem the world developed was data science. The data science problem means that data is growing at a much faster rate than computational solution techniques, and it is the result of the fact that storage cost is getting cheaper gradually day by day, therefore, keeping data safe and secure for further use becomes cheaper with time. Social activities, biological explorations, scientific experiments, along with the sensor devices are great data contributors. Data science is beneficial to the society and business but at the same time, it brings challenges to the scientific communities. The existing traditional tools, machine learning algorithms and techniques are not capable of handling, managing and analyzing big data. Although various scalable machine learning algorithms, techniques and tools are prevalent. In this paper we have identified the most relevant issues and challenges related to data science and point out a comprehensive comparison of various technical Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education; Computer Science- Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI; Mathematics - Mathematical Modeling in various scientific methods like Statistics, Statistical and Stochastic modeling, Probability and queuing for handling data science problem.

III. DATA SCIENCE AND ACADEMICS

In the words of Alex Szalay, these sorts of researchers must be "Pi-shaped" as opposed to the more traditional "T-shaped" researcher. In Szalay's view, a classic PhD program generates T-shaped researchers: scientists with widebut-shallow general knowledge, but deep skill and expertise in one particular area. The new breed of scientific researchers, the data scientists, must be Pishaped: that is, they maintain the same wide breadth, but push deeper both in their own subject area and in the statistical or computational methods that help drive modern research.

IV. DATA MINING/SCIENCE WITH BIG DATA

Aspects of big data have been studied and considered by a number of data mining researchers over the past decade and beyond. Mining massive data by scalable algorithms leveraging parallel and distributed architectures has been a focus topic of numerous workshops and conferences. However, the embrace of the Volume aspect of data is coming to a realization now, largely through the rapid availability of datasets that exceed terabytes and now penta bytes—whether through scientific simulations and experiments, business transactional data or digital footprints of individuals. Astronomy, for example, is a fantastic application of big data driven by the advances in the astronomical instruments. Each pixel captured by the new instruments can have a few thousand attributes and translate quickly to a penta scale problem. This rapid growth in data is creating a new field called Astro-informatics, which is forging partnerships between computer scientists, statisticians and astronomers. The emergence of big data from various domains, whether in business or science or humanities or engineering, is presenting novel challenges in scale and provenance of data, requiring a new rigor and interest among the data mining community to translate their algorithms and frameworks for data-driven discoveries.

V. ESSENTIAL POINTS

Big Data has given rise to Data Science; Data science is rooted in solid foundations of mathematics and statistics, computer science, and domain knowledge, not everything with data or science is Data Science, The use cases for Data Science are compelling.

VI. DATA VARIETY

Data presents itself in varied forms for a given concept. It is presenting a new notion to learning systems and computational intelligence algorithm for classification, where the feature vector is multi-modal, with structured and unstructured text, and still the notion is to classify one concept from another. How do we create a feature vector, and then a learning algorithm with an appropriate objective function to learn from such varied data?

A. Data size

On one hand, we develop "one-pass learning" algorithms that require only one scan of the data with limited storage irrelevant to data size; on the other hand, we try to identify smaller partitions of the really valuable data from the original big data.

B. Data trust

While data is rapidly and increasingly available, it is also important to consider the data source and if the data can be trusted. More data is not necessarily correct data, and more data is not necessarily valuable data. A keen filter for the data is a key.

VII. DISTRIBUTED EXISTENCE

Owners of different parts of the data might warrant different access rights. We must aim to leverage data sources without access to the whole data, and exploit them without transporting the data. Extreme distribution: Taking this idea even further, the unit-level data may be what we see as the level of data distribution, as we deal with issues of privacy and security. New approaches to modeling big data will be required to work with extreme distributed data.

VIII. DIVERSE DEMANDS

People may have diverse demands whereas the high cost of big data processing may disable construction of a separate model for each demand. Can we build one model to satisfy the various demands? We also need to note that, with big data, it is possible to find supporting evidence to any argument we want; then, how to judge/evaluate our "findings"?

IX. SUB-MODELS

Diverse demands might also relate to diversity of the behaviors that we are modeling within our application domains. Rather than one single model to cover it all, the model will consist of ensembles of a large number of smaller models that together deliver better understandings and predictions than the single, complex model.

X. INTUITION IMPORTANCE

Data is going to power novel discoveries and action oriented business insights. It is important to still attach intuition, curiosity and domain knowledge without which one may become myopic and fall in the chasm of "correlation is enough". Computational intelligence should be tied with human intuition.

XI. RAPID MODEL

As the world continues to "speed up", decisions need to be made more quickly because fraudsters can more quickly find new methods in an agile environment, model building must become more agile and real-time. Algorithms such as meta-heuristics have achieved great success in academic research, but have rarely been employed in industry. One major obstacle is the huge computational cost required for evaluating the quality of candidate designs of complex engineering systems. The emerging big data analytic technologies will remove the obstacle to a certain degree by reusing knowledge extracted from the huge amount of high-dimensional, heterogeneous and noisy data. Such knowledge can also be acquired with new visualization techniques. Big data driven optimization will also play a key role in reconstruction of large-scale biological systems.

XII. COMPLEX OPTIMIZATION

Definition of decision variables, setup of the objectives and articulation of the constraints are three main steps in formulating optimization problems before solving them. For optimization of complex systems, formulation of the optimization problem itself becomes a complex optimization problem. The big data approach might provide us new insights and methodologies for formulating optimization problems, thus leading to a more efficient solution. In closing the discussion, we emphasize that the opportunities and challenges brought by big data are very broad and diverse, and it is clear that no single technique can meet all demands. In this sense, big data also brings a chance of "big combination" of techniques and of research.

XIII. CONCLUSION

Data science is a very new in application, and this field has much more importance in daily life.

REFERENCES

[1] J. Abello, P. M. Pardalos, and M. G. Resende, Handbook of Massive Data Sets, ser. Massive Computing. Springer, 2002, vol.

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

[3] C. C. Aggarwal, Data streams: Models and algorithms. Springer, 2007, vol. 31.

[4] Australian Department of Immigration, "Fact sheet 70 - managing the border," Internet, 2013. [Online]. Available: http://www.immi.gov.au/media/factsheets/70border.htm

[5] J. Bader and E. Zitzler, "HypE: An algorithm for fast hyper volume-based many objective optimization," Evolutionary Computation, vol. 19, no. 1, pp. 45– 76, 2011.

[6] P. Baldi and P.J. Sadowski, "Understanding drop out, "in Advances in Neural Information Processing Systems26. 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 Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 2001, pp. 26–33.

[8] A.-L. Barab´asi, Linked: The New Science Of Networks. Basic Books, 2002.

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

[10] Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning,vol.2,no.1,pp. 1–127, 2009.

[11] E. Brynjolfsson, L. Hitt, and H. Kim, "Strength in numbers: How does data-driven decision making affect firm performance?" Available at SSRN 1819486, 2011.

[12] T. Chai, Y. Jin, and S. Bernhard, "Evolutionary complex engineering optimization: Opportunities and challenges," IEEE Computational Intelligence Magazine, vol. 8, no. 3, pp. 12–15, 2013.

[13] N. V. Chawla, "Data mining for imbalanced datasets: Anover view,"in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. US: Springer, 2005, pp. 853–867.

[14] N. V. Chawla et al., Learning on Extremes-Size and Imbalance-of Data. Florida, US: University of South Florida, 2002.

[15] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with augmented class by exploiting unlabeled data," in Proceedings of the 28th AAAI Conference on Artificial Intelligence, Quebec City, Canada, 2014.

[16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182–197, 2002.

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

[18] D. Easley and J. Kleinberg, Networks, crowds, and markets. Cambridge Univ Press, 2010.

[19] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," Pattern Recognition, 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 IEEE Congress on 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, "One-pass AUC optimization," in Proceedings of the 30th International Conference on 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, 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 IEEE Congress on Evolutionary Computation, Cancun, Mexico, 2013, pp. 1491–1498.

[28] H. Ishibuchi and T. Murata, "Multi-objective genetic local search algorithm," in Proceedings of IEEE International Conference on Evolutionary Computation, Nagoya, Japan, 1996, pp. 119–124.