A Review of Bayesian Statistical Methods Applied to Big Data

Farzana Jahan[1] and Kerrie Mengersen[2]

[1]PhD Student, Science and Engineering Faculty, Queensland University of Technology

[2] Distinguished Professor of Statistics, Science and Engineering Faculty, Queensland University of Technology

Introduction

Sheer amount of information is now being generated, accumulated, collected or integrated in every fraction of second formulating the term "Big Data". There has been no unique definition of Big Data, to date there is a large body of literature which attempt to define Big Data effectively. (1,2)

A Definition of Big Data

"Big Data includes the following aspects: "Volume", "Velocity" and "Variety", to describe the characteristics of information, "Technology" and "Analytical Methods", to describe the requirements needed to make proper use of such information, and "Value", to describe the transformation of information into insights that may create economic value for companies and the society." (1)

There have been a wide range of research going on involving Big Data. We can broadly categorize the published literature on Big Data as follows:

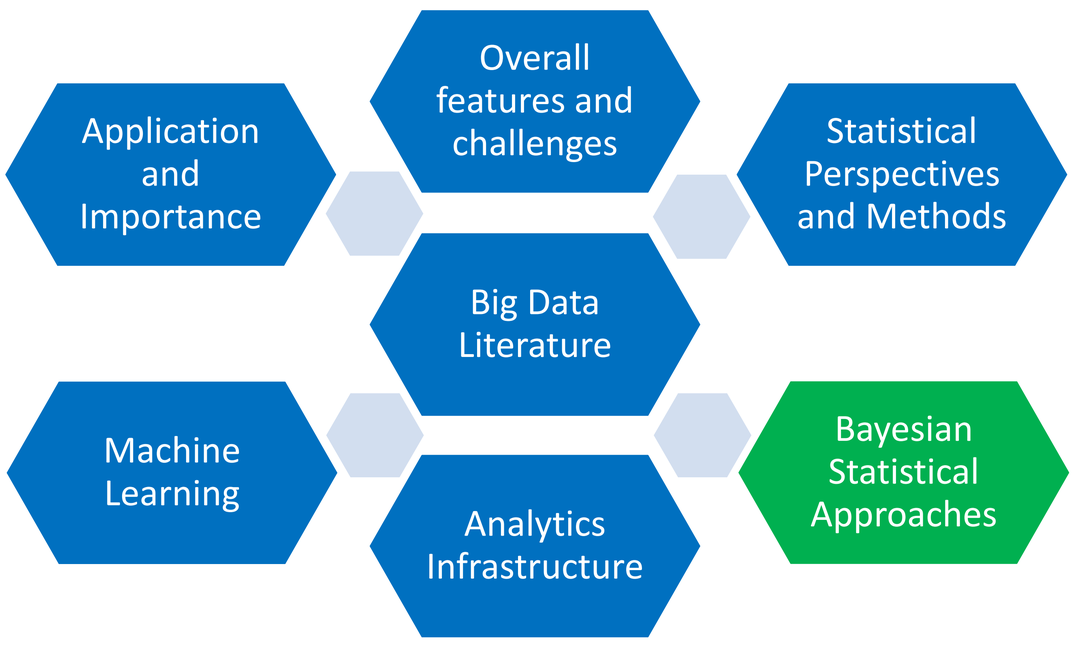

Fig 1: Classification of Big Data Literature

Objectives of the Study

- This study attempts to review the published studies that present Bayesian statistical models specifically for Big Data and discuss the reported and perceived benefits of these approaches.
- We aim to answer, whether focusing only on improving computational algorithms and infrastructure will be enough to face the challenges of Big Data.

Motivation of the Study

- Although, there have been many review papers on Big Data (1,3,4,5,6,7 etc.), to the best of our knowledge, to the best of out knowledge, the survey of Bayesian Statistical modelling applied to Big Data to address our question above remains unexplored.
- We attempt to make a potential contribution to the knowledgebase for the Statisticians to understand the usefulness and necessity to develop scalable Bayesian Statistical modelling for Big Data.

Bayesian Approaches in Big Data

The number of articles applying Bayesian approaches to Big Data is not very large (9,10,11,12)

Majority of the works conducted from the perspectives of Bayesian statistics were concerned with designing scalable algorithms to be able to analyse Big Data as noted from the reviews made. There is lack of research on Bayesian Modelling in Big Data situations.

Figure 2: A Word Cloud of most frequent terms prepared from the review of literature involving Bayesian techniques in Big Data using NVivo 10.

Critical Reflection

- Though, we can see emphasis of theoretical developments regarding new methods or models to analyse Big Data in some articles (13,14,15), little work is concentrated on statistical or in particular Bayesian Statistical modelling to Big Data.
- The advantages of Bayesian Statistical modelling as: incorporation of multiple sources of information into the model via prior and flexibility in model estimation with high dimensional parameter space etc. (16) which are still under explored in Big Data perspectives.

Conclusion

- We are living in the era of Big Data and continuous research works are in progress to make the best use of the available data. Our review identified the application or development of Bayesian statistical modelling of Big Data.
- This review identified the need for further research to identify the advantages and/or drawbacks of Bayesian Statistical modelling in Big Data situations. The area is yet to be explored to be able to answer whether existing models are enough to face the Big Data challenges with scaling algorithms or there is need for new models to be developed for improved decision making with Big Data.
- As literature suggests, careful modelling and sound theoretical platform is needed along with the computational advances (13,14,15), we may conclude that there should be more works concentrating on Bayesian Statistical modelling in Big Data perspectives in order to reveal the usefulness of Bayesian inference and also to identify the challenges to address while developing or applying Bayesian Statistical models for Big Data.

Key References

1. De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features. Library Review. 2016;65(3):122-35.
2. Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics. 2015;165:234-46.
3. Bhosale HS, Gadekar DP. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications. 2014;4(10):1-7.
4. Emani KC, Cullot N, Nicolle C. Understandable Big Data: A survey. Computer Science Review. 2015;17:70-81.
6. Fan J, Han F, Liu H. Challenges of Big Data Analysis. Natl Sci Rev. 2014;1(2):293-314.
7. Khan N, Yaqoob I, Hashem IA, Inayat Z, Ali WK, Alam M, et al. Big data: survey, technologies, opportunities, and challenges. ScientificWorldJournal. 2014;2014:1-18.
8. Labrinidis A, Jagadish HV, editors. Challenges and Opportunities with Big Data. VLDB Endowment; 2012.
9. Allenby GM, Bradlow ET, George EI, Liechty J, McCulloch RE. Perspectives on Bayesian Methods and Big Data. Customer Needs and Solutions. 2014;1(3):169-75.
10. Balakrishnan S, Madigan D. A One-Pass Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets. Bayesian Analysis 2006;1(2):345-62.
11. Liu B, Blasch E, Chen Y, Shen D, Chen G, editors. Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier. 2013 IEEE International Conference on Big Data; 2013.
12. Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE. Bayes and big data: the consensus Monte Carlo algorithm. International Journal of Management Science and Engineering Management. 2016;11(2):78-88.
13. Franke B, Plante J-F, Roscher R, Lee E-sA, Smyth C, Hatefi A, et al. Statistical Inference, Learning and Models in Big Data. International Statistical Review. 2016;84(3):371-89.
14. Hilbert M. Big Data for Development: A Review of Promises and Challenges. Development Policy Review. 2015;martinhilbert. net. Retrieved:10-07.
15. Wise AF, Schaffer DW. Why theory matters more than ever in the age of big data. Journal of Learning Analytics. 2015;2(2):5-13.
16. Congdon P. Bayesian statistical modelling: John Wiley & Sons; 2007.

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS