International Journal of Research
Available at https://edupediapublications.org/journals
Special Issue on Conference Papers
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 06
March 2018

Review on "Data Mining with Big Data"

K Krishnaiah Mr.L.Kiran Kumar Reddy
Assistant Professor – CSE, Visvesvaraya College of Engineering & Technology
krishna.koneti16@gmail.com kirankumarreddy_1976@yahoo.co.in

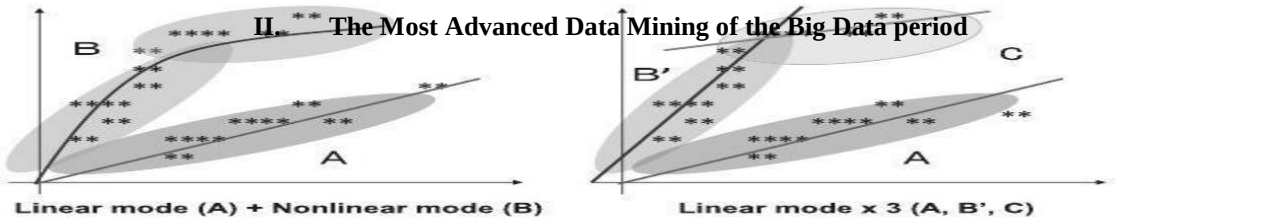
Abstract—Big Data relates large-volume, complex, increasing data sets with multiple independent sources. With the rapid evolution of data, data storage and the networking collection capability, Big Data are now speedily expanding in all science and engineering domains. Big Data mining is the ability of extracting constructive information from huge streams of data or datasets, that due to its variability, volume, and velocity. Data mining includes exploring and analyzing big quantity of data to locate different molds for big data. Artificial intelligence (AI) and statistics are the fields which develop these techniques, This paper discusses a characterizes applications of Big Data processing model and Big Data revolution, from the data mining outlook. The analysis of big data can be troublesome because it often involves the collection and storage of mixed data based on different patterns or rules (heterogeneous mixture data). This has made the heterogeneous mixture property of data a very important issue. This paper introduces —heterogeneous mixture learning,‖ We study the tough issues in the Big Data revolution and also in the data-driven model.

Index Terms — Big Data; data mining; heterogeneous mixture; autonomous sources; complex and evolving associations

Introduction

With the exponential development of data comes an ever-growing requirement to route and evaluate the so-called Big Data. Heavy performance computing structures have been devised to attend the needs for managing Big Data methods not only from an operation processing point of view but also from an analytics view. The most important target of this paper is to offer the reader with a historical and complete view on the current style in the direction of huge performance computing architectures specially it transmit to Data Mining and Analytics .There are a series of readings discretely on Big Data (and its individuality), High presentation Computing for Massively Parallel Processing (MPP) databases, Analytics and algorithms for Big Data. In-memory Databases, implementation of mechanism learning algorithms for Big Data proposals, the Analytics environments of the future, etc. though none gives a chronological and broad vision of all these split topics in a particular document. It is the author's first try to bring about as several of these topics mutually as probable and to describe an ideal analytic environment that is superior to the challen ges of today's analytics requirement. Modern production trends advise that big data investigation is becoming necessary for involuntary discovering of intelligence that is concerned in the repeatedly-occurring patterns and unseen rules. These may then be used efficiently as helpful information (such knowledge-inventing technology is usually referred to as data mining). For example, electricity demand is predicted by extracting the convention leading the values of a range of sensors such as thermometers and of electricity demand and deriving future demand predictions by applying such rules to the current sensor data. In this paper, we first discuss the difficulties of heterogeneous mixture data analysis. In short, the impossibility of per-forming exhaustive searches due to the huge number of data grouping candidates, which in reality symbolizes the essential difficulty of the analysis. Next, we introduce heterogeneous mixture learning. This is the most advanced heterogeneous data analysis technology to be developed at NEC. It features the application of an advanced machine learning technology called the —factorized asymptotic Bayesian inference,‖ and we will focus mainly on the introduction of its fundamental concept. Finally, we introduce a demonstration experiment of electricity demand prediction for a building as an example of a suitable application of heterogeneous mixture learning. With the heterogeneous mixture learning technology, we have succeeded in improving the prediction Big Data skills has broadly classified in three tasks: Data Analysis, Development and Big Data Infrastructure. Software Development abilities can be auxiliary divided transversely domains such as Big Data- Database, Big Data-Development. Data Analysis includes two domains: Data Mining Statistical Analysis and BI & Visualization Tools.

II. The Most Advanced Data Mining of the Big Data period



For the proper analysis of data, it is required to find the optimum grouping method from a large set of data grouping candidates. (The ellipses correspond to the data grouping methods and the lines to the prediction models.)

Fig. 1 Illustration of heterogeneous mixture data

accuracy by 7.6 points (10.3% → 2.7%) compared to the previous prediction method without considering the heterogeneous mixtur e data, and by 2.1 points (4.8% → 2.7%) compared to the method that is dependent on data grouping by experts.

A. Issue of Heterogeneous mixture data analysis

One of the key points in the accurate analysis of heterogeneous mixture data is to break up the inherent heterogeneous mixture properties by arranging the data in groups having the same patterns or rules. However, since there are a huge number of possibilities (sometimes infinite) for the data grouping options, it is in reality impossible to verify each and every candidate. The following three issues are of importance in arranging the data into several groups.

1) Number of groups (How much the data is mixed)
2) Method of grouping (How the data is grouped)
3) Appropriate choice of prediction model according to the properties of each group

These issues cannot be solved independently or by following the order from 1) to 3), but they should be solved simultaneously by considering their mutual dependences. For example, when the hypothesis is that data contains a mixture of nonlinear and linear relationships (Fig. 1, Left), a highly accurate prediction model can be obtained by grouping the data into two groups (ellipse B and ellipse A). However, when the hypothesis is that the data contains a mixture of multiple linear relationships (Fig. 1, Right), the optimum number of groups becomes 3. In both left and right parts of Fig. 1, the grouping methods (ellipses) are determined by the sets of data to which the linear (or nonlinear) relationships (prediction models) are applicable, and this fact means that it is not possible to determine 2) by ignoring 1) and 3).

It is obligatory then to consider issues 1) to 3) simultaneously, which is the specific number of data grouping candidates. As an example, let us assume a case in which big data storage of a large volume of sensor and electricity demand data is analyzed to detect the hidden rules. Furthermore, to clarify the essence of this issue, we will limit the candidates for the prediction model (electricity demand prediction formula) to those that can be expressed by a quadratic expression of the explanatory variables (sensor values). When the number of explanatory variables (number of sensors) is fixed at 10, the number of sensors usable in the prediction model at 3 and the number of groups obtained by data grouping at 4, the number of prediction model candidates is calculated approximately at(100 C 3) 4 = 6.84×10 20 (10 20 is equal to 1 trillion multiplied by 100 millions). In more complicated cases, there are almost infinite combination candidates of data groups and prediction models. This means that the time taken for a search is at an unrealistic level if simple algorithms are used. As described in section 1, the solution most often adopted hitherto to solve such a problem was to define the factors altering the rules via trial and error based on expert knowledge and to classify the data accordingly in order to enable the automatic extraction of a single rule for each group. However, to determine the optimum data grouping method for data acquired from such a complex system is very difficult to achieve, even for experts. Constraints are posed by a reduction in the prediction accuracy due to inappropriate grouping and by the huge amount of labor required for the trial and error procedures needed to find the optimum grouping method.

B. Data mining based on heterogeneous mixture learning

NEC has developed a new heterogeneous mixture learning technology for use in mining heterogeneous mixture data. This technology is capable of the high speed optimization of the three issues 1) to 3) referred to in section 2 above by avoiding issues related to data grouping or a sudden increase in prediction model combinations Below, we explain the differences between learning with the previous techniques (such as the cross-validation or the Bayesian information criterion) and the heterogeneous mixture learning as shown in Fig. 2. Previous techniques calculated the scores (information criteria) for the model candidates and selected the model with the best score. However, as we described in section 2 above, an unrealistic calculation time would be required if these techniques were applied to the learning of heterogeneous mixture data due to the enormous number of model candidates. On the other hand, heterogeneous mixture learning is capable of adaptive searching of issues 1) to 3), which are the number of groups, the method of grouping and the prediction model for each group. This makes it possible to find the optimum data grouping and prediction model by investigating models with high prediction accuracies without searching unpromising candidates. The advanced search and optimization of the heterogeneous mixture learning is backed by the latest machine learning theory called ―factorized asymptotic Bayesian inference‖ 2)3)4) .

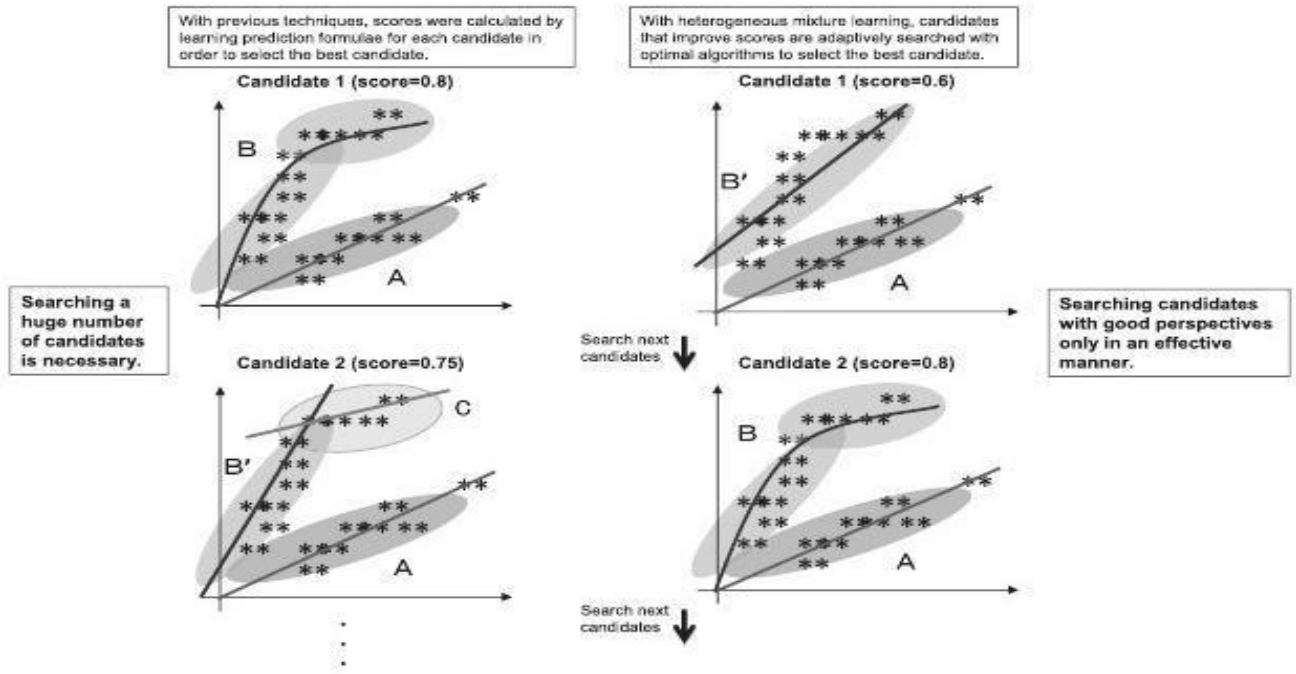


Fig. 2 Differences in data grouping and prediction model search methods between heterogeneous mixture learning and previous t echniques**.**

III. Big Data Means

Big data is classically described by the first three properties below—occasionally referred to as the three but organizations require a fourth value to build big data job

A. **Volume:** massive information sets that are command of size bigger than data managed in habitual storage and analytical results. Imagine petabytes rather than terabytes.

B. **Variety:** complex, variable and Heterogeneous data, which are generated in formats as dissimilar as public media, e-mail, images ,video, blogs, and sensor data—as well as ―shadow data‖ such as access journals and Web explore histories.

C. **Velocity:** Data is generated as a stable with real-time queries for significant information to be present up on claim instead of batched.

D. **Value:** consequential insights that transport predictive analytics for upcoming trends and patterns from bottomless, difficult analysis based on graph algorithms, machine learning and statistical modeling. These analytics overtake the results of usual querying, reporting and business intelligence.

IV. Data Mining for Big Data

Data mining includes extracting and analyzing bulky amounts of data to discover models for big data. The methods came out of the grounds of artificial intelligence (AI) and statistics with a tad of database management.

Searching information from data takes two major forms: prediction and description. it is tough to know what the data shows?. Data mining is used to summarize and simplify the data in a way that we can recognize and then permit us to gather things about specific cases based on the patterns

Normally, the objective of the data mining is either prediction or classification. In classification, the thought is to arrange data into sets. For example, a seller might be attracted in the features of those who answered versus who didn't answered to a advertising.

There are two divisions. In prediction, the plan is to predict the rate of a continuous variable. For example, a seller might be involved in predicting those who *will* reply to a promotion.

Distinctive algorithms used in data mining are as follows:

A. **Classification trees:** A famous data-mining system that is used to categorize a needy categorical variable based on size of one or many predictor variables. The outcome is a tree with links and nodes between the nodes that can be interpret to form if-then rules.

B. **Logistic regression:** A algebraic technique that is a modification of standard regression but enlarges the idea to deal with sorting. It constructs a formula that predicts the possibility of the occurrence as a role of the independent variables.

C. **Neural networks:** A software algorithm that is molded after the matching architecture of animal minds. The network includes of output nodes, hidden layers and input nodes. Each unit is allocated a weight. Data is specified to the input node, and by a method of trial and error, the algorithm correct the weights until it reaches a definite stopping criteria. Some groups have likened this to a black–box system.

D. **Clustering techniques like K-nearest neighbors:** A procedure that identifies class of related records. The K-nearest neighbor technique evaluates the distances between the points and record in the historical data. It then allocates this record to the set of its nearest neighbor in a data group.

CONCLUSION

Big data is directed to continue rising during the next year and every data scientist will have to handle a large amount of data every year .This data will be more miscellaneous, bigger and faster. We discussed in this paper several insights about the subjects and what we think are the major concern and the core challenges for the future. Big Data is becoming the latest final border for precise data research and for business applications. Data mining with big data will assist us to discover facts that nobody has discovered before. The heterogeneous mixture learning technology is an advanced technology used in big data analysis. In the above, we introduced difficulties that are inherent in heterogeneous mixture data analysis, the basic concept of heterogeneous mixture learning and the results of a demonstration experiment that deal with electricity demand predictions. As the big data analysis increases its importance, heterogeneous mixture data mining technology is also expected to play a significant role in the market. The range of application of heterogeneous mixture learning will be expanded broader than ever in the future. To investigate Big Data, we have examined a number of challenges at the system levels, data and model. To hold Big Data mining, high-performance computing platforms are necessary, which enforce organized designs to set free .the complete power of the Big Data. By the data level, the independent information sources and the range of the data gathering environments, habitually result in data with complex conditions, such as missing unsure values. The vital challenge is that a Big Data mining structure needs to consider complicated interaction between data sources ,samples and models along with their developing changes with time and additional probable factors. A system wants to be cautiously designed so that unstructured data can be connected through their composite relationships to form valuable patterns, and the development of data volumes and relationships should help patterns to guess the tendency and future.

References

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, senior Member,IEEE,Gong-Qing,Wu,and Wei Ding, senior Member,IEEE:Data Mining with big Data
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014
[2] M.H. Alam, J.W. Ha, and S.K. Lee, ―Novel Approaches to Crawling Important Pages Early,‖ Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, ―Identifying Influential and Susceptible Members of Social Networks,‖ Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, ―Big Privacy: Protecting Confidentiality in Big Data,‖ ACM Crossroads, vol. 19, no. 1, pp.20-
23, 2012.
[5] FUJIMAKI Ryohei, MORINAGA Satoshi :The Most Advanced Data Mining of the Big Data Era
[6] E. Birney, ―The Making of ENCODE: Lessons for Big-Data Projects,‖ Nature, vol. 489, pp. 49-51, 2012.
[7] J. Bollen, H. Mao, and X. Zeng, ―Twitter Mood Predicts the Stock Market,‖ J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, ―Network Analysis in the Social Sciences,‖ Science, vol. 323, pp. 892-895, 2009.
[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.
[10] D. Centola, ―The Spread of Behavior in an Online Social Network Experiment,‖ Science, vol. 329, pp. 1194-1197, 2010.

[11] E.Y. Chang, H. Bai, and K. Zhu, ―Parallel Algorithms for Mining Large-Scale Rich-Media Data,‖ Proc. 17th ACM Int'l Conf. Multi-media, (MM '09,) pp. 917-918, 2009.
[12] R. Chen, K. Sivakumar, and H. Kargupta, ―Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,‖
Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.
[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, ―Efficient Algorithms for Influence Maximization in Social Networks,‖ Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.
[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, ―Map-Reduce for Machine Learning on Multicore,‖
Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.
[15] G. Cormode and D. Srivastava, ―Anonymized Data: Generation, Models, Usage,‖ Proc. ACM SIGMOD Int'l Conf. Management Data,
pp.1015-1018, 2009.
[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, ―Ricardo: Integrating R and Hadoop,‖ Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998. 2010.
[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, ―The Square Kilometre Array,‖ Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.
[18] P. Domingos and G. Hulten, ―Mining High-Speed Data Streams,‖
Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.
[19] G. Duncan, ―Privacy by Design,‖ Science, vol. 317, pp. 1178-1179, 2007.
[20] B. Efron, ―Missing Data, Imputation, and the Bootstrap,‖ J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.
[21] A. Ghoting and E. Pednault, ―Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics,‖ Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.