AN ANALYSIS OF BIG DATA DIMENSIONALITY REDUCTION TECHNIQUE

Prem Kumar Chandrakar, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India, (premchandrakar@gmail.com) A. K. Shrivas, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India, '(akhilesh.mca29@gmail.com)

ABSTRACT

The big data refers to data index whose volume (size), variability (complexness), and rate of velocity (growth) shuffling them complicated to be captured, managed, process or analyze. Research on big data examination is approaching in the new stage called rapid data where distinctive gigabytes of data meet up in the big data structures reliably. Present big data structures assemble typically complex data streams in view of the big data offering to 6Vs like volume, velocity, value, variety, variability, and veracity. After the reducing of big data a resultant data to be more valuable as contrast with collecting raw, redundant, inconsistent, and noisy data. A further point of view for reducing big data is that large data sets of millions of variables cause the dimension curse that requires unlimited IT resources to discover patterns of process able knowledge. This paper presents a review of methods that are used for big data reduction and conjointly presents an in depth discussion on big data dimension reduction methods, big data compression, size reduction, redundancy elimination, data extraction, and automatic learning methods.

Keywords: Big Data, Data Compression, Data Reduction, Data Complexity, Dimensionality Reduction.

INTRODUCTION

The word big data describe inventive technique and technology to collect, store up, allocate, handle and examine petabyte or bigger size of data sets among towering rapidity and dissimilar structure. System generated big data might be prearranged (structured), formless (unstructured) or semi-structured, so the data resultant in helplessness of usual data management methods. Data sets of big data are create from various sources and can enter in the system at a valid range. To process these immense measures of data in a modest and proficient way, parallelism is utilized. The big data refers to data sets whose volume (size), variability (complexity), and rate of velocity (growth) make them complicated and not to be easy for capturing, managing, processing or analyzing by commonly used standard technologies , like database software such as relational database and statistical analysis or visualization software , in the moment needed to compose them of use (Ur Rehman et al., 2016).

Big data is that the grouping of huge scale, voluminous, and multi data format originated from various sources (Che, Safran, & Peng, 2013). Volume is the main quality of big data to be exact instead of by the in front of storage spaces in large-scale data centers and storage area networks. The enormous size of the big information causes the information heterogeneous as well as results in various dimensionalities in the datasets. Hence, Hence, endeavors are required to diminish the volume to adequately break down enormous information (Che et al., 2013). One more, big data stream are should have been prepared online to stay away from auxiliary asset utilization for capacity and handling. The second key normal for big data velocity. The frequency of data streams is a velocity, which is looked-for to be reduced in order to handle big data efficiently. For example, solar powered flow observatory creates a one terabytes data for each day and the investigation of such a quick enormous data is conceivable simply after reduction (Battams, 2015). Conversely big data inherit the curse of dimensionality or millions of dimensions (variables, features, attributes) are necessary to be well reduced to find out the maximum knowledge patterns (Zhai, Ong, & Tsang, 2014) (Fan, Han, & Liu, 2014). For example, individual genomic high-throughput sequencing not simply grows the volume and speed of data yet furthermore adds to the high dimensionality of the data (Chandramouli, Goldstein, & Duan, 2012)(Ward, Schmieder, Highnam, & Mittelman, 2013). Therefore, it is fundamental to lessen the high measurements while holding the most imperative and valuable information.

This paper presents an exhaustive writing survey of techniques for huge big data reduction. A couple of comparable earlier investigations have additionally been directed. In any case, these examinations either exhibit a non specific exchange of big data diminishment or a specific game plan of critical structures or techniques. Additionally, various open research issues are likewise introduced to coordinate future researchers. The primary commitments of this paper are:

i. A complete literature review and classification of big data reduction method acting are presented.

- ii. In current earned run average planned schemes for big data decrease are analyzed and synthesized.
- iii. A complete gap analysis for the voice of constraints and future research challenge for information reduction in big data environments is presented.

The composition is organized as different section like: In the second section we can discuss what is the importance of big data reduction and discusses the complexity of big data and next section presented various big data reduction methods. Section number fourth presented the open issues and future research challenges and lastly, the paper is concluded.

COMPLEXITY OF BIG DATA AND WHY WE NEED FOR DATA REDUCTION

The big data collections contain different arrangements of online networking information, modern sensor systems, logical trial frameworks, associated welling, and a few other application regions. The data gathered from large scale nearby and remote detecting gadgets and systems, Internet-empowered information streams, as well as gadgets, frameworks, and systems logs brings greatly heterogeneous, multi-source, multi-organize, collected, and persistent big data streams. Successfully taking care of the big data stream to store, file, and question the information hotspots for sidelong information handling is among the key difficulties being tended to by analysts (Hashem et al., 2015) (Gani, Siddiqa, Shamshirband, & Hanum, 2016)(Gani et al., 2016). Be that as it may, data researchers are confronting information storm issue to reveal the most extreme learning designs at the fine-grained level for powerful and customized use of big data frameworks (Battams, 2015) (Kambatla, Kollias, Kumar, & Grama, 2014). The data storm is because of 6Vs properties of big data, specifically the volume, variety, value, velocity, veracity, and variability. The authors (Gani et al., 2016) discussed about the 6Vs.

- i. *Volume-* The size of data characterizes as a volume of big data. Be that as it may, there is no endless supply of big data which indicates the amount of data to be considered as big keeping in mind the end goal to meet the meaning of big data. However, any data size as big regarding volume which isn't effortlessly handled capable by basic computing systems.
- ii. *Velocity* The big data speed is controlled by the recurrence of information streams which are entering in big data systems. The velocity is taken care of by big data frameworks in two ways. First, the entire data streams are gathered in concentrated systems, and afterward, encourage data processing is performed.
- iii. *Variety-* The big data system catching data stream from different data sources which create data streams in multiple formats. In this manner, big data system must have the capacity to process numerous sorts of the data stream so as to adequately uncover hidden knowledge patterns.
- iv. *Veracity* The value of big data systems increments when the data streams are gathered from solid sources. The data stream gathering is performed with compromising the quality of data streams.
- v. *Variability* Since all data sources in big data frameworks don't create the information streams with a similar speed and same quality. Along these lines, variability property empowers to deal with the significant issues.
- vi. *Value-* The esteem property of big data characterizes the utility, ease of use, and value of big data system. This property tends more toward the results of data analytics and data processing processes and is specifically corresponding to different 5Vs in huge information frameworks.

The big data frameworks must competent to manage every one of the 6Vs proficiently by making a harmony between information preparing destinations and the cost of information handling in big data systems. Furthermore, the complexity in big data systems develops in three structures: (1) information many-sided quality, (2) computational multifaceted nature, and (3) framework unpredictability (Jin, Wah, Cheng, & Wang, 2015). The data complexity quality emerges because of various arrangements and unstructured nature of huge information, which hoist the issue of different measurements and the complex entomb dimensional and intra-dimensional connections. For instance, the semantic connection between various estimations of a similar quality, for instance, the commotion level in the specific territories of the city, builds the between dimensional unpredictability.

Big data are otherwise called huge information mining is a tedious errand interfacing superfluous endeavors to lessen information to a reasonable size to reveal most extreme learning designs. To make it useful for information investigation, various pre-handling methods for outline, drawing, peculiarity identification, measurement diminishment, commotion expulsion, and exception's recognition are connected to lessen, refine, and clean big data (Li & Nath, 2014). The New York Times, a main US daily paper, reports that information researchers invest 50–80% of the energy in cleaning the enormous data (Lohr S ,2014). The terms utilized as a part of the business for the previously mentioned process are information robbing, information wrangling, or information janitor work. Another issue with the expansive scale high dimensional information examination is the more right of taking in models that are created from huge quantities of characteristics with a couple of cases. These learning models fit well inside the

preparation information, yet their execution with testing information altogether corrupts (Ma, Zhang, & Wang, 2014).

Data association is one all the more part to talk about the big data lessening issue. The viable huge information administration assumes a critical part from information securing to examination and representation. Despite the fact that information securing from different sources and total of important datasets enhance the productivity of big data frameworks, it builds the in-organize preparing and information development at bunches and server farm levels. So also, the ordering systems talked about in (Gani et al., 2016) upgrade the huge information administration; be that as it may, the strategies go over information preparing overheads. In spite of the fact that the transformation of unstructured information to semi-organized and organized configurations is valuable for successful inquiry execution, the change in itself is a period and asset devouring movement. In addition, big data is immense in volume that is circulated in various storerooms. Subsequently, the improvement of learning models and revealing worldwide information from enormously conveyed huge information is a repetitive undertaking.

As of late, organizations especially the endeavors are transforming into big data system. The accumulation of big data streams from Web clients' close to home information streams (click-streams, ambulation exercises, geo-areas, and wellbeing records) and mix of those information streams with customized administrations is a key test (Ur Rehman et al., 2016). The gathering of unessential data streams builds the computational weight that straightforwardly influences the operational cost of ventures. In this way, the social occasion of fine-grained, exceptionally pertinent, and decreased data spill out of clients is another test that requires genuine consideration while planning big data system. Right now, client data gathering by outsiders without express assent and data about commercialization is raising the protection issues.

The center innovative help for big data reduction techniques depends on multilayer design. The data stockpiling is empowered by substantial scale server farms and systems of various figuring bunches (Majeed and Shah, 2015). The capacity foundations are overseen by center systems administration administrations, embarrassingly parallel dispersed processing structures, for example, Hadoop map-reduce implementations and large- scale virtualization technologies (Ahmad et al., 2015) (Bonomi, Milito, Natarajan, and Zhu, 2014)(Ur Rehman, Liew, and Wah, 2014). At the most reduced layers of the huge data, design lives the multi-organize information sources which incorporate large-scale logical data streams among numerous others. This layered engineering empowers to process and oversee enormous information at numerous levels utilizing different registering frameworks with various shape factors. Subsequently, wide scopes of use models are outlined and new system have been created for big data handling.

REDUCTION METHODS OF BIG DATA

In this part we show the data reduction strategies being connected in big data systems. Utilizing big data decrease techniques improve the capacity of data or reduce data redundancy and duplication. A portion of the strategies just reduce the volume by compacting the first data and a portion of the techniques lessen the speed of information streams at the most punctual before entering in huge information stockpiling frameworks. On the other hand, a portion of the strategies extricate topological structures of unstructured information and diminish the general huge information utilizing system hypothesis approaches that are talked about as takes after (Ur Rehman et al., 2014).

FEATURE SELECTION OF BIG DATA

Highlight Feature selection is a system which is utilized to locate the great nature of significant highlights from the first dataset utilizing some goal measures. These days, Feature Selections have turned out to be testing issues in the field of Pattern Recognition (Kumar and Elavarasan, 2014) Machine Learning (Crone and Kourentzes, 2010), Data Mining (Smialowski, Frishman, and Kramer, 2009) and Case-Based Reasoning (Balamurugan and Rajaram, 2009). Highlight Selection is a procedure of finding an ideal or imperfect subset of x highlights from the first X highlights. It requires an extensive large space to get the ideal component subset. The ideal element subset is measured by assessment criteria. The fundamental goal of the element choice is to decrease the quantity of highlights and to evacuate the superfluous, repetitive and boisterous information (Balamurugan and Rajaram, 2009). By lessening the highlights, one can decrease the framework multifaceted nature, over fitting of learning strategies and increment the computational speed.

FEATURE SELECTION PROCEDURE OF BIG DATA

The big data highlight feature selection procedure is fundamental four stages; subset age, subset assessment, stopping criterion and result validation (Dong, Hua, and Li, 2007).

i. Subset Generation

It is a big procedure that produces the applicant includes subset utilizing certain inquiry technique. The procedure has two fundamental issues. They are, seek course and inquiry procedure. Right off the bat, a beginning stage must be chosen which thusly impacts the inquiry heading. The search strategies are categorized into complete search, sequential search an random search (Dong et al., 2007).

ii. Subset Evaluation

The subset assessment, it is the assessment standard is utilized to assess each recently created subset. The assessment standard is utilized to decide the integrity of the subset. The assessment criteria are isolated into Independent, Dependent and Hybrid criteria (Coelho, Braga, and Verleysen, 2010).

iii. Stopping Criteria

It is utilized to stop the element determination process. The component choice process may stop under one of the accompanying criteria (Dong et al., 2007). A predefined number of highlights is chosen, A predefined number of emphases is come to, on the off chance that, expansion (or cancellation) of a component neglects to deliver a superior subset, an ideal subset as per the assessment standard is gotten.

iv. Validation

The validation procedure is utilized to measure the resultant subset utilizing the earlier information about the information. In a few applications, the pertinent highlights are known in advance, an examination is done between the known arrangements of highlights with the chose highlights.

FEATURE SELECTION APPROACHES OF BIG DATA

There are the various feature selection approaches of big data such as filter approach, wrapper approach, hybrid approach (Ramaswami and Bhaskaran, 2009). Some of statistical measures are used to finding the feature selection such as Information Theory, Mutual Information, Information Gain (IG), Gain Ratio (GR), Symmetric Uncertainty (SU), Correlation- Based Feature Selection (Zhang, 2008).

FEATURE EXTRACTION OF BIG DATA

The component extraction system is utilized to get the most important data from the first information and speak to that data in a lower dimensionality space. This procedure is utilized to choose another arrangement of highlights. The component change might be a direct or nonlinear blend of unique highlights. The highlights are extricated utilizing the accompanying techniques (H. (National U. of S. Liu, Motoda, Setiono, & Zhao, 2010).

Principal Component Analysis (PCA) (Liu, Motoda, Setiono, & Zhao, 2010) is a classical statistical technique which is widely used to reduce the dimensionality of a dataset consisting of enormous amount of interrelated variables. PCA reduces the dimensionality by transforming the original dataset into a new set of variables, called principal components, where the largest variance present in the original dataset is captured with the maximum component in order to extract the most important information.

OPEN RESEARCH ISSUES OF BIG DATA

Study of the big data diminishment is performed at many levels amid the data preparing lifecycle that incorporate information catching, information preprocessing, information ordering and capacity, information investigation, and representation. In this manner, the pertinent decrease strategies and frameworks ought to be intended to deal with the big data multifaceted nature at all phases of big data handling. Another future research work should concentrate on considering each of the 6Vs to process big data in figuring frameworks with various shape factors from fine-grained portable registering systems to large-scale massively parallel computing infrastructures (Gani et al., 2016).

CONCLUSION

Big data complexity is a key issue that is needed to be mitigated. The methods discussed in this paper are an effort to address the issue. All of the presented literature review reveals that there is no existing method that can handle the issue of big data complexity single handedly by considering the all 6Vs of big data. The studies discussed in this

paper mainly focused on data reduction in terms of volume (by reducing size) and variety (by reducing number of features or dimensions). On the other hand, further efforts are required to reduce the big data streams in terms of velocity and veracity. In addition, the new methods are required to reduce big data streams at the earliest immediately after data creation and its entrance into the big data frameworks. As a rule, pressure based information diminishment strategies are helpful for decreasing volume. On the other hand, the decompression overhead should be considered to enhance effectiveness.

In this paper, a study has been carried out to know how the high dimensionality issue has been unraveled utilizing distinctive dimensionality lessening techniques. In this paper we are attempt to gives entire information about how the huge information highlights was chosen and removed utilizing highlight choice and highlight extraction techniques respectively. In feature selection technique the most relevant features are selected using statistical measure and some of the statistical measure are explained in detail. In feature extraction technique the new feature were obtained from the original features using the various statistical techniques and most popular statistical techniques were explained in detail. Hence this paper will help the beginners who were doing research in the dimensionality reduction techniques.

REFERENCES

- Ahmad, R. W., Gani, A., Hamid, S. H. A., Shiraz, M., Yousafzai, A., & Xia, F. (2015). A survey on virtual machine migration and server consolidation frameworks for cloud data centers. Journal of Network and Computer Applications, 52, 11–25.
- Balamurugan, S. A. A., & Rajaram, R. (2009). Effective and efficient feature selection for large-scale data using Bayes' theorem. *International Journal of Automation and Computing*, 6(1), 62–71.
- Battams, K. (2015). Stream mining for solar physics: Applications and implications for big solar data. *Proceedings* 2014 IEEE International Conference on Big Data, IEEE Big Data 2014, 18–26.
- Bonomi, F., Milito, R., Natarajan, P., & Zhu, J. (2014). Big Data and Internet of Things: A Roadmap for Smart Environments, 546, 169–186.
- Chandramouli, B., Goldstein, J., & Duan, S. (2012). Temporal analytics on big data for web advertising. *Proceedings - International Conference on Data Engineering*, 90–101.
- Coelho, F., Braga, A. P., & Verleysen, M. (2010). Selection and Model Selection Based on Pearson's Correlation Coefficient. Proceedings of the 2012 International Conference of Modern Computer Science and Applications, 509–516.
- Crone, S. F., & Kourentzes, N. (2010). Feature selection for time series prediction A combined filter and wrapper approach for neural networks. *Neurocomputing*, 73(10–12), 1923–1936.
- Dong, M., Hua, J., & Li, Y. (2007). A Gaussian Mixture Model to Detect Clusters Embedded in Feature Subspace. Communications in Information and Systems, 7(4), 337–352.
- Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data Analysis. Natl Sci Rev, 1(2), 293-314.
- For Big-Data Scientists, 'Janitor Work' to Is Key Hurdle to Insights. (Browsing Date: 8-11-1017).
- Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: taxonomy and performance evaluation. *Knowledge and Information Systems*, 46(2), 241–284.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.
- Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. Big Data Research, 2(2), 59–64.
- Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7), 2561–2573.
- Kumar, V. A., & Elavarasan, N. (2014). A Survey on Dimensionality Reduction Technique. International Journal of Emerging Trends & Technology in Computer Science, 3(6), 36–41.
- Li, F., & Nath, S. (2014). Scalable data summarization on big data. *Distributed and Parallel Databases*, 32(3), 313–314.
- Liu, H. National U. of S., Motoda, H. Osaka U., Setiono, R., & Zhao, Z. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining. *Journal of Machine Learning Research: Workshop and Conference Proceedings 10: The Fourth Workshop on Feature Selection in Data Mining*, 4–13.
- Ma, C., Zhang, H. H., & Wang, X. (2014). Machine learning for Big Data analytics in plants. *Trends in Plant Science*, 19(12), 798-808.
- Majeed, A., & Shah, M. A. (2015). Energy efficiency in big data complex systems: a comprehensive survey of modern energy saving techniques. *Complex Adaptive Systems Modeling*, 3(6), 1-29.

- Ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big Data Reduction Methods: A Survey. *Data Science and Engineering*, 1(4), 265–284.
- Ur Rehman, M., Liew, C. S., & Wah, T. Y. (2014). UniMiner: Towards a unified framework for data mining. Information and Communication Technologies (WICT), 2014 Fourth World Congress on, 134–139.
- Ward, R. M., Schmieder, R., Highnam, G., & Mittelman, D. (2013). Big data challenges and opportunities in high-throughput sequencing. *Systems Biomedicine*, 1(1), 29–34.
- Xia, W., Jiang, H., Feng, D., & Hua, Y. (2011). Silo: a similarity-locality based near-exact deduplication scheme with low ram overhead and high throughput. *Proceedings of The 2011 USENIX Annual Technical Conference (USENIX ATC '11)*, 26–28.
- Zhang, Z. (2008). Mining relational data from text: From strictly supervised to weakly supervised learning. Information Systems, 33(3), 300–314.