Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.114 – 116

SURVEY ARTICLE

CLUSTERING HIGH DIMENSIONAL BIG DATA- A LITERATURE SURVEY

S. Vydehi MCA., M.Phil.[1], **R. Suganya** M.Sc (CT)., [2]

[1]Head & Assistant Professor, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India

[2]M.Phil Scholar (Computer Science), Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu, India

[1] vydehiela@gmail.com; [2] rsuganya.2292@gmail.com

***Abstract*— In modern business for data transfer into an informational advantage is an increasing important tool seen by data mining. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The most prominent data mining technique is clustering for grouping data into clusters based on its distance measures. The challenging fact is clustering big data that is huge in their dimensions and needed numerous resources for its process. The process of grouping into high dimensional data into clusters is not accurate and perhaps not up to the level of expectation when the dimension of the dataset is high. The tremendous attention towards clustering a big data is the recent trend among the researchers. In this literature survey, analysis of various clustering algorithms considering the criteria of big dataset and on high dimensional dataset how clustering is done.**

***Keywords*— Data Mining, clustering, k-means, k-medoids, Artificial intelligence, BAT and big data**

I. INTRODUCTION

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web. With the beginning in the era of big data, the data is increasing at rapid speed not only in size but also in variety. There comes challenge and difficulties to handle such large amount of data with the growing data. [1]Big data exhibits different characteristics like volume, variety, variability, value, velocity and complexity due to which it is very difficult to analyse data and obtain information with traditional data mining techniques. Data mining is the process of analysing data from different context and encapsulating it into useful information. Data mining consists of extracting, transforming, and loading transactional data into the data warehouse system, storing and managing the data in a multidimensional database system, providing data access to business analysts and information technology professionals, analysing the data by application software, presenting the data in a useful format .Data mining includes the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, analysis of various clustering algorithms considering the criteria of big dataset and on high dimensional dataset how clustering is done

II. RELATED WORK

High Performance Multidimensional Scaling

[2] Seung-Hee Bae, Judy Qiu has proposed High Performance Multidimensional Scaling for Large High-Dimensional Data Visualization in their work they have described a well-known dimension reduction

algorithm, called MDS (SMACOF),and have discussed how to utilize the algorithm for a huge data set. The main issues involved in dealing with a large amount of data points are not only lots of computations but also huge memory requirements. Parallelization via the traditional MPI approach in order to utilize the distributed memory computing system, which can support much more computing power and extend the accessible memory size, is proposed as a solution for the amendment of the computation and memory shortage so as to be able to treat large data with SMACOF. They have also discussed for performance analysis for maximizing the performance of parallelization the process of data composition structure is highly considered.

Application of real data set

[3] Tian Zhang, Raghu Ramakrishnan, Miron Livny in their survey related to clustering large real data set paper, an efficient and scalable data clustering method is proposed, based on a new in-memory data structure called CF-tree, which serves as an in-memory summary of the data distribution. they have implemented it in a system called BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and studied its performance extensively in terms of memory requirements, running time, clustering quality, stability and scalability; we also compare it with other available methods. Finally, BIRCH is applied to solve two real-life problems: one is building an iterative and interactive pixel classification tool, and the other is generating the initial codebook for image compression.

Towards Ultrahigh Dimensional Feature Selection for Big Data

[4] Mingkui Tan, Ivor W. Tsang, Li Wang in their work a new adaptive feature scaling scheme for ultrahigh-dimensional feature selection on Big Data, and then reformulate it as a convex semi-infinite programming (SIP) problem. To address the SIP, they propose an efficient feature generating paradigm. Different from traditional gradient-based approaches that conduct optimization on all input features, the proposed paradigm iteratively activates a group of features, and solves a sequence of multiple kernel learning (MKL) sub problems. To further speed up the training, we propose to solve the MKL sub problems in their primal forms through a modified accelerated proximal gradient approach. Due to such optimization scheme, some efficient cache techniques are also developed. The feature generating paradigm is guaranteed to converge globally under mild conditions, and can achieve lower feature selection bias. Moreover, the proposed method can tackle two challenging tasks in feature selection: 1) group-based feature selection with complex structures, and 2) nonlinear feature selection with explicit feature mappings. Comprehensive experiments on a wide range of synthetic and real-world data sets of tens of millions data points with O(1014) features demonstrate the competitive performance of the proposed method over state-of-the-art feature selection methods in terms of generalization performance and training affiance.

Mining High Dimensional Data Sets Using Big Data

[5] G. Yogaraj, A. Arumuga Arun have discussed challenge and different techniques to handle large dataset like big data To support Big Data mining, high performance computing platforms are required which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors.

Statistical Analysis of Big Data

[6] Jianqing Fan and Han Liu in their work they discusses statistical methods for estimating complex correlation structure from large pharmacogenomic datasets. They selectively overview several prominent statistical methods for estimating large covariance matrix for understanding correlation structure, inverse covariance matrix for network modeling, large-scale simultaneous tests for selecting significantly differently expressed genes and proteins and genetic markers for complex diseases, and high dimensional variable selection for identify important molecules for understanding molecule mechanisms in pharmacogenomics. Their applications to gene network estimation and biomarker selection are used to illustrate the methodological power. Several new challenges of Big data analysis, including complex data distribution, missing data, measurement error, spurious correlation, endogeneity, and the need for robust statistical methods, are also discussed.

Sparse Representations of High Dimensional Data

[7] Zhen James Xiang Hao Xu Peter J. Ramadge has proposed a discussion of Learning sparse representations on data adaptive dictionaries are a state-of-the-art method for modeling data. But when the dictionary is large and the data dimension is high, it is a computationally challenging problem. They explore three aspects of the problem. First, they derive new, greatly improved screening tests that quickly identify code words that are guaranteed to have zero weights. Second, they studied the properties of random projections in the context of learning sparse representations. Finally, they developed a hierarchical framework that uses incremental random projections and screening to learn, in small stages, a hierarchically structured dictionary for sparse representations. Empirical results show that their framework can learn informative hierarchical sparse representations more efficiently.

III.FUTURE WORK

So far various ways of dealing and handling large dimensional dataset has considered the Incremental Affinity Propagation Clustering Based on Message Passing [8] by Leilei Sun and Chonghui Guo is existing system they have proposed an algorithm IAP clustering based on K-Medoids (IAPKM) and IAP clustering based on Nearest Neighbor Assignment (IAPNA) they have checked with five popular dataset and in proposed new system KMBAT algorithm will be generated that is combination of K-Medoids clustering algorithm and BAT an artificial intelligence algorithm uses echo system as its base. Dataset of social network will take and connection of HADOOP will be established for best clustering result for big data.

IV. CONCLUSIONS

While the concept of incremental queries has been known for some time, the clustering implications have not been explored with users. In particular, it has been an open question whether data analysts would be comfortable interacting with confidence intervals. We hope that showing the utility of these approximations will encourage further research on both the front- and back-ends of these systems. Hope new algorithm of KMBAT will explore all the disadvantages of clustering high dimensional big data because it will use progressive fetch mechanism so clustering will be considered easy.

REFERENCES

[1] Avita Katal Mohammad Wazid R H Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", 2013 IEEE
[2] Seung-Hee Bae, Judy Qiu, "High Performance Multidimensional Scaling for Large High-Dimensional Data Visualization", 2012 IEEE
[3] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: A New Data Clustering Algorithm and Its Applications", springer.
[4] Mingkui Tan, Ivor W. Tsang, Li Wang, "Towards Ultrahigh Dimensional Feature Selection for Big Data", 2014 Journal of Machine Learning Research.
[5] G. Yogaraj, A. Arumuga Arun, " Mining High Dimensional Data Sets Using Big Data", 2015 International Journal of Advanced Research in Computer Science and Software engineering.
[6] Jianqing Fan and Han Liu, "Statistical Analysis of Big Data on Pharmacogenomics",2013.
[7] Zhen James Xiang Hao Xu Peter J. Ramadge, "Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries".
[8] Leilei Sun and Chonghui Guo, "Incremental Affinity Propagation Clustering Based on Message Passing", 2014 IEEE.

