Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

Enhanced ECLAT algorithm for frequent item set mining using ECLAT- A review

Bharati Suvalka
M.Tech. (C.S.E.)
Geetanjali Institute of Technical Studies
Udaipur Rajasthan Technical University, Kota, Rajasthan, India

Astract: presenting an review paper on big data, their analytics using machine learning and then enhancing ECLAT algorithm for frequent item set mining using revised ECLAT**.**

INTRODUCTION

Sagiroglu et al. [1] explain the big data content, scope, methods, samples, advantages and confront of Data. The vital issue about the Big data is the seclusion and protection. Big data samples describe the make another study of about the research. By this paper, we can conclude that any organization having big data can get the advantage from its careful analysis for the problem solving function. Using Knowledge innovation from the Big data is easy to get the information from the complex data sets. The overall valuation describe that the data is ever-increasing and becoming complex. The confront is not only to collect and manage the data also how to take out the useful information from that collected facts. According to the Intel IT Center, there are lots of challenges related to Big Data which are data growth, data infrastructure, data multiplicity, data revelation, data rate.

Mukherjee A. et al [2] explains that Big data analytics describe the analysis of large amount of data to obtain the useful information and uncover the unseen patterns. Big data analytics refers to the Map reduce Framework which is developed by the Google. Apache Hadoop is the open source platform which is used for the reason of operation of Google's map Reduce model.

Garlasu et al [3] offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing provides the concept of distributed computing. The assistance of Grid computing center is the high storage ability and the high processing power. Grid Computing makes the big offerings among the scientific research, help the scientists to analyze and store the large and intricate data.

Aditya B. Patel et al. [4] done the experimental work on the Big data problems. It illustrate the best solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage space and Map Reduce programming framework for parallel processing to progression of large data set.

Dr. Siddaraju [5] explained various methods for catering to the problems in hand through MapReduce framework over HDFS. MapReduce technique has been studied at in this paper which is needed for implementing Big Data analysis using HDFS.

Lior Rokach et al. [6] reports method and criterion that are used for determining whether two objects are similar or dissimilar. Then the clustering methods are presented, divided into: hierarchical, partitioning, density-based, model-based, grid-based, and soft-computing methods. Consequent the methods, the challenges of performing clustering in large data sets.

Bikash Sharma et al. [7] discussed case for a hybrid data center consisting of native and virtual environments, and propose a 2-phase hierarchical scheduler, called HybridMR, for the effective resource management of interactive and batch workloads. In the first phase, HybridMR classifies incoming MapReduce jobs based on the accepted virtualization overheads, and uses this information to automatically guide placement between physical and virtual machines. In the second phase, HybridMR

manages the runtime performance of MapReduce jobs collocated with interactive applications in order to provide best effort delivery to batch jobs, while complying with the Service Level Agreements (SLAs) of interactive applications.

Hui Gao et al. [8] presents Text clustering is one of the hard and hot research fields in the text mining research. Combining Map Reduce framework and the neuron initialization method of VPSOM (vector pressing Self Organizing Model) algorithm, a new clustering algorithm is presented. It divides the large text vector dataset into data blocks, each of which then processed in dissimilar distributed data node of Map Reduce framework hierarchical clustering algorithm.

Tilmann Rabl et al. [9] describes complete performance evaluation of six modern (open-source) data stores in the perspective of application performance monitoring as part of CA Technologies inventiveness. They evaluate these systems with data and workloads that can be found in application presentation monitoring, as well as, on-line advertisement, power monitoring, and other use cases.

M.JAYASREE [10] explained about fast growth of networks these days organizations have filled with the collection of millions of data with large number of combinations. This big data challenges over business problems. It requires more analysis for the high-performance process. The new methods of Hadoop and MapReduce methods are discussed from the data mining perspective.

Ricardo B. C. Et al. [11] presented the use of Ranking Meta-Learning approaches to ranking and selecting algorithms for problems of time series forecasting and clustering of gene expression data. Given a problem of forecasting or clustering, the Meta-Learning approach provides a ranking of the candidate algorithms, according to the features of the problem's dataset. The best ranked algorithm can be returned as the particular one. In order to evaluate the Ranking Meta-Learning proposal, prototypes were implemented to rank artificial neural networks models for forecasting financial and economic time series and to rank clustering algorithms in the context of cancer gene expression datasets.

Amresh Kumar et al. [12] confirmed and validate various MapReduce applications like wordcount, grep, terasort and parallel K-Means Clustering Algorithm. It has been initiate that as the number of nodes increases the completing time decreases, but also some of the interesting cases has been found during the experiment and recorded the various performance change and drawn different performance graphs.

Jeffrey Dean and Sanjay Ghemawat [13] present implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable a typical MapReduce calculation processes many terabytes of data on thousands of machines. Programmers and the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day.

Aristides Gionis et al [14] consider the follow problem: given a set of clustering, find a clustering that agrees as much as feasible with the given clustering. This problem, clustering aggregation, appears as expected in various contexts. For example, clustering definite data is an instance of the problem: each definite variable can be viewed as a clustering of the input rows. Moreover, clustering aggregation can be used as a meta-clustering method to improve the strength of clustering. The problem formulation does not require a prior information about the number of clusters, and it gives a normal way for handling mislaid values. They give a formal statement of the clustering aggregation problem.

Vadivel.M , Raghunath.V [15] This paper introduces our skill of grouping internet users by mining a huge volume of web access log of up to 500 gigabytes. The purpose is realized using hierarchical clustering algorithms with Map-Reduce, a parallel processing framework over clusters. However, the immediate execution of the algorithms suffers from efficiency problem for both insufficient memory and higher execution time. This paper presents an proficient hierarchical clustering method of mining large datasets with Map-Reduce. The method include two optimization techniques: Batch Updating to reduce the

computational time and communication costs with cluster nodes, and Co-occurrence based feature selection to decrease the dimension of feature vectors and remove noise features.

Ella Peltonen [16] This paper has introduced the current fields of Big Data, cloud computing and distributed machine learning. It has presented what cloud computing environments can provide for analyzing large data sets efficiently. Especially, they presented Berkeley Data Analysis Stack and two of its systems: the dynamic cluster resource manager, and the in-memory cluster computing system Spark. Spark extends a well-known computing paradigm MapReduce. In contrast to MapReduce that provides two functions only, map and reduce, the Spark system offers a diverse library of functions and memory control operators.

O. L. Mangasarian [17] This paper describes four developments, one theoretical, three algorithmic, all centered on support vector machines (SVMs). SVMs have become the tool of choice for the fundamental classification problem of machine learning and data mining.

Stan Salvador and Philip Chan [18] propose an proficient algorithm, the L method, that find the "knee" in a '# of clusters vs. clustering evaluation metric' graph. Using the knee is well-known, but is not a predominantly well-understood method to resolve the number of clusters. They search the feasibility of this method, and try to determine in which situations it will and will not work. Also compare the L method to existing methods based on the accuracy of the number of clusters that are resolute and effective. There results show approving performance for these criteria compared to the existing methods that were evaluated.

Anne Denton et al. [19] In this paper a hierarchical method is introduce that is fundamentally related to partitioning methods, such as k-medoids and k-means, as well as to a density based method, namely center-defined DENCLUE. It is greater to both k-means and k-medoids in its reduction of outlier pressure. Nevertheless it avoids both the time complexity of some partition-based algorithms and the storage requirements of density-based ones. An implementation is presented that is particularly suited to spatial, stream, and multimedia data, using P-trees for proficient data storage and access.

Timothy C. Havens et al. [20] In this paper, they propose an extension to the scalable VAT(sVAT) algorithm that produces crisp c-partitions of big data. Our algorithm first uses sVAT to sample the data and reorder the sample. We then use a property of VAT reordering that allows us to efficiently compute single-linkage partitions of the reordered data sample. Finally, we expand the partition to the entire data set using the nearest object rule. This also describes the sVAT algorithm, followed by our projected sVAT-SL algorithm. Demonstrate the proposed sVAT-SL on synthetic and real data.

Limou Wang [22] This paper present SVM and complete description of Hierarchical clustering algorithm and Clustering-Based SVM (CB-SVM).

Hwanjo Yu et al. [23] This paper presents a new method, Clustering-Based SVM (CB-SVM), which is exclusively designed for managing very large data sets. CB-SVM applies a hierarchical micro-clustering algorithm that scans the entire data set only once to offer an SVM with high quality samples that carry the statistical summaries of the data such that the summaries maximize the advantage of learning the SVM. CB-SVM tries to generate the best SVM boundary for very large data sets given limited amount of resources.

REFERENCES

1. Sagiroglu, S.; Sinanc, D. (20-24 May 2013),"Big Data: A Review" in IEEE International Conference on Computer Science and Electronics Engineering, 2013.
2. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. (18-22 Dec. 2012) "Shared disk big data analytics with Apache Hadoop" in IEEE International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 2013.
3. [3] Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. ;(17-19 Jan. 2013),"A Big Data implementation based on Grid Computing" in the third International Conference on Communications and Information Technology ICCIT 2013.

4. Aditya B. Patel, Manashvi Birla, Ushma Nair (6-8 Dec. 2012) "Addressing Big Data Problem Using Hadoop and Map Reduce" in IEEE International Conference on Data Mining and Network Technologies , 2012.
5. Dr. Siddaraju1, Sowmya C, L Rashmi K, Rahul M " Efficient Analysis of Big Data Using Map Reduce Framework " in International Journal of Computer Science and Technology IJCST Vol.2, Issue 2, 2011.
6. Lior Rokach, Oded Maimon"CLUSTERING METHODS" in IEEE 2nd International Conference on data mining and Networking , 2013.
7. Bikash Sharma, Timothy Wood, Chita R. Das HybridMR: A Hierarchical MapReduce Scheduler for Hybrid Data Centers" in International Conference on Emerging Big data Web Technologies, 2013.
8. Hui Gao, Jun Jiang, Li She, Yan Fu "A New Agglomerative Hierarchical Clustering Algorithm Implementation based on the Map Reduce Framework" " in the International Conference on Communications and Information Technology ICCIT 2012.
9. Tilmann Rabl, Mohammad Sadoghi, HansArno Jacobsen, Victor Munt´esMulero"Solving Big Data Challenges for Enterprise Application Performance Management" in Workshop on Hot Topics in Cloud Computing (Hot Cloud), San Diego, USA, 2009.
10. M.JAYASREE " Data Mining: Exploring Big Data Using Hadoop and MapReduce" in International Conference on Computer Science and Electronics Engineering, 2012.
11. Ricardo B. C. Prudˆencio, Marcilio C. P. de Souto, and Teresa B. Ludermir "Selecting Machine Learning Algorithms Using the Ranking Meta-Learning Approach" in International Journal of Computer Science and Technology IJCST Vol.4, Issue 2, 2011.
12. Amresh Kumar, Kiran M. , Ravi Prakash G.,Saikat Mukherjee " Verification and Validation of MapReduce Program model for Parallel K-Means algorithm on Hadoop Cluster " in IEEE International Conference on Communication Systems and Network Technologies (CSNT), 2012. and Conservation of Energy (ICGCE), 2013.
13. Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplifed Data Processing on Large Clusters" in IEEE International Conference on Computer Science and Electronics Engineering, 2012.
14. Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas "Clustering Aggregation" in International Conference on Communication Systems and data mining, 2011.
15. Vadivel.M , Raghunath.V "Enhancing Map-Reduce Framework for Bigdata with Hierarchical Clustering" in Workshop on Software Engineering Challenges of big data, Vancouver, Canada, May 2009.
16. Ella Peltonen "An approach to Machine Learning with Big Data" O. L. Mangasarian "Data Mining via Support Vector Machines" in IEEE International Conference on Big data and Network Technologies, 2011.
17. Stan Salvador and Philip Chan "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms" in IEEE International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 2013.
18. Anne Denton, Qiang Ding, William Perrizo Qin Ding "Efficient Hierarchical Clustering of Large Data Sets Using P-trees" in the International Conference on Data Mining and Information Technology 2013.
19. Timothy C. Havens , James C. Bezdek , Marimuthu Palaniswami "Scalable Single Linkage Hierarchical Clustering For Big Data" in International Journal of Computer Science and Technology IJCST Vol.2, Issue 2, 2012.
20. Limou Wang "Classifying Large Data Sets Using SVMs with Hierarchical Clusters" in IEEE International Conference on Computer Science and Electronics Engineering, 2013.
21. Hwanjo Yu, Jiong Yang, Jiawei Han "Classifying Large Data Sets Using SVMs with Hierarchical Clusters " in the International Conference on Communications and Information Technology ICCIT 2012.

