Comparative study of conventional Pattern Mining Algorithms on Internet of Things

[1]Neha Kumari, [2]Monika Saxena, [3]Dr.C.K.Jha
[1]Research Scholar, [2]Assistant Professor, [3]Proffesor (HOD)
[1]Masters of Technology,
[1]Banasthali vidyapeeth, Jaipur, India

Abstract : Pattern mining algorithms is used to mine the useful data from the massive amount of IOT data. Mostly used data mining algorithms are classification, clustering, association rule and regression in which the classification and regression comes under supervised learning and other two in unsupervised learning. The objective is to review different techniques applied for mining the pattern by using classification and clustering algorithms. We have discussed the advantages and Dis-advantages of different algorithms coming under classification and clustering. By applying parallel data mining algorithm in map reduce frame work.

IndexTerms - data mining, classification, clustering, map reduce, parallel data mining

I. INTRODUCTION

Data mining is the techniques developed to handle wide amount of data using different tools to process it properly. It involves discovering interesting and useful pattern from large data sets and to extract the hidden information we applies the algorithms. The main reason behind the development of wide amount of data is the availability of information and increase in processing power. Fig 1 illustrates the process of data mining. Data preparation in which data is firstly prepared for mining then integrate the data from various sources, clean the noise and extract the hidden information. Apply algorithm to evaluate pattern, at last the knowledge is then represented to the user [1].

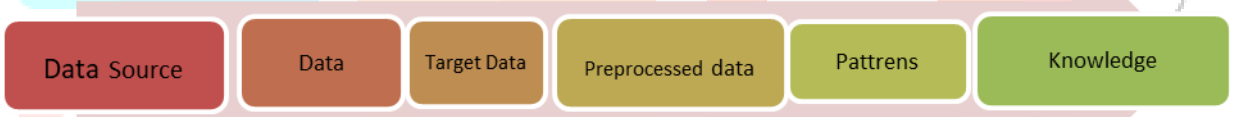

FIG 1: Data mining overview

The data mining models for the internet of things were discussed by shen bin et al. [2].
Technologies can continuously integrate classical networks with network instrument and devices. IoT brings great challenge in order to maintain and analyze the data for future use.[3].

I.I PARALLEL AND DISTRIBUTED DATA MINNING

Sujni Paul [4] has describes the method of parallelism. Problems like memory and CPU speed limitations faced by the single processors. So to solve these problems we have parallelism algorithms. There are two approaches:
Task parallel algorithm: the work of this algorithm is to assign the portion of search space to the single processor. It is divided into two groups .first group is divide and conquer that separate the search space and allocation of each portion to the specific processor. And second is Task queue that dynamically assign the smaller portion of search space to the processor when it become available.
Data parallel algorithm: Distribute the available data to the processors that are free for allocation. Data parallel algorithms have two ways. Record based partition that assigns non overlapping sets of record to each processor. And assigning sets of attributes to each processor is attribute based portioning

I.II MAP REDUCES MODEL:

Map reduce is the framework in which we can write applications to process large amount of data parallelly. It consists of two functions that are Map and reduce. The mapper processes the data and creates several small chunks of data. The Reducer's job is to process the data that comes from the mapper. The map reduce model is proposed by the Google and it is an open source .Map reduce is the framework in which we can write applications to process large amount of data parallelly. The data has given the records in the list forms that are represented in the form of (key and value). The key and values are in the pairs. It consists of two functions that are Map and reduce. The mapper processes the data and creates several small chunks of data. The Reducer's job is to process the data that comes from the mapper. The results were stored in the file system that is known as distributed file system. The map reduce model have lots of advantages such as, fault tolerant, easily expandable. If the one node of the system gets

affected by the problem then there will be no effect on the second node of the system, they were just have to process the data of that node again on the other node.

I.III SUPERVISED AND UNSUPERVISED LEARNING:

Supervised learning is the process in which output datasets is provided which is used to train the machine and then apply that knowledge to the test data. As per the result it gets the desired output. Classification and regression comes under supervised learning

Unsupervised learning is the process in which instead of providing output datasets to train the machine, machine learns through observation and find structure in data. Clustering and association rule comes under unsupervised learning.

II. CLASSIFICATION:

Classification is the task of finding a set of models or functions that define and apart data classes. For management of decision making classification is very important. To precisely predict the target class for each case in data is the goal of classification.. Suppose we have an object, then assigning it to one of the predefined target is classification . Finding a set of functions or models that define and separate data classes is the job of classification. Classification is very important for the management of decision making.

James m Keller et al [5] describes the fuzzy sets into the K nearest neighbor technique to develop the fuzzy version of the algorithm. They presented the three levels of assigning fuzzy membership to the particular sample. They describe that not only the fuzzy algorithm dominates its counter parts in terms of lower error rate but the confidence measure of the classification is given by resulting membership.

Matthew et al [6] presents the decision tree algorithm in serial implementation. Some serial decision tree algorithms are ID3, C4.5, SLIQ, SPRINT, CART etc. the decision tree classification is done in two phases: The tree building and pruning. Serial implementation is fast, memory resident and easy to understand as compared to parallel.

Ming yang el at [7] presents the Bayesian network learning model to study the data mining techniques. They firstly established the Bayesian network learning model then the parameters of reorganization and at last the selections of coefficients were analyzed. There is the deduction in the computation of Bayesian, then the reliability of the model is verified by the example.

Anurag et al [8] presents the parallel decision tree algorithm based on induction. They described two basic parallel formulations, one is synchronous tree construction approach and another one is partitioned based construction approach. They describes the advantages and disadvantages of each method, they proposed one hybrid method from above two methods that gives the best features in order to reduce communication overhead and load imbalance.

. **Gongqing et al [9]** presents the map reduce c4.5 algorithm that is the new method which combines the advantages of C4.5 with the map reduce computing model with implementing ensemble learning. They were build on clusters and it was constructed only once but used anywhere. Some issues are there such as effectiveness of this algorithm on large clusters.

Qing et al [10] presents the parallel implementation of KNN, decision tree, naïve Bayesian model on the large datasets and have the result that it has the property of linear scalability. We have to improve the usages efficiency of computing resources

Zhu Fubao et al [16] has proposed the cart decision tree algorithm based on map reduce attribute weights. They have proposed this algorithm in order to remove the previous problems like low efficiency, memory consumption, bad accuracy, and complexity. The algorithms work on map reduce model. The comparison table illustrate that it is very important to marks the attributes by weight through map reduce. It improves the efficiency of the spanning tree by marking the attributes by weight through map reduce model. The accuracy of the modified algorithm is also improved slightly as compared to the previous one but still very low. As per the practical result they concluded that the accuracy of decision table created by weight in classification is more than the other algorithms.

Batra Mridula et al [17] has presented the comparative study of decision tree algorithms that are ID3 (iterative Dichotomizer 3), CART (Classification and regression tree), CHAID (chi-squared automatic interaction detector), C4.5. The motto of the paper is to describe the detail analysis of decision tree algorithms and its variation in order to determine the best decision for the purpose of predication. As per result they concluded that ID3 does not provide the best solution and doesn't provide the optimal least feasible tree solution. In C4.5 algorithm if the attributes values are not known then they are not used in information gain. CART algorithm is used to efficient tool in order to release the data relationship which remains hidden using other data analytical tools. CHAID provide us the multi way frequency tables that are very useful for decision making purpose.

Reference no:	Title of the paper	Methodology/ Algorithm	Advantages	Disadvantages	Work on map reduce
5	A Fuzzy K-Nearest Neighbour Algorithm	KNN into fuzzy sets	It is attractive and produces membership assignments that are desirable.	fuzzy nearest prototype classifier, has more error rates as compared to fuzzy nearest neighbour classifier,	no
6	Comparative Analysis of Serial Decision Tree Classification Algorithms	Serial implementation of decision tree algorithm	Easy and cheap to implement	Slow than the parallel implementation	no
7	Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm	Bayes network algorithm	Capable for handling large amounts of complex data in real applications, it can use its reasoning and self-learning ability.	the determination of the prior density , Bayes network requires a variety of assumptions for the premise, it has no existing rules,	no
8	Parallel formulation of decision tree classification algorithm	Parallel decision tree algorithm. Dynamic load balancing with processor groups to reduce processor idling. 25	Dynamic load balancing with processor groups to reduce processor idling.	Difficult to parallelize due to the inherent dynamic nature of the computation.	yes
9	MReC4.5: C4.5 ensemble classification with Map Reduce	Map reduce C4.5 algorithm	Easy to implement and improve the effectiveness.	Not enough good for large Clusters.	Yes
10	Parallel implementation of classification algorithm based on	parallel KNN, Decision tree, Naïve Bayesian	Process large data .and scalability	Improvement in usages of resources	Yes

III.CLUSTERING:

Clustering is known as the process of gathering same data into the groups. The clustering algorithm is used to cluster the large datasets into the small no of many groups having similarity in nature. And should be dissimilar to the objects in the different groups. Here we will have different types of clustering algorithm used for grouping of data into the groups.

Manish et al [11] presents the k-means algorithm which partition n observations into the K clusters, here k is the no of clusters we want for the experiment. For K mean clustering firstly randomly select the clusters, calculate the distance between the data points and cluster centers. Assign the point to the minimum distance cluster. Recalculate the distance of new center with points. So on. They concluded that it has fastest and better performance as compared to other algorithms, only problem is that it is sensitive for noise in large data sets.

Lokeshwari et al [12] present the parallel kmean algorithm that takes advantages of data parallelism in manner to equally divide the load work to all the processors. Data objects are divided into small and similar size blocks, and then each block is assigned to processor. Every processor do the work of computing and comparing the centroid of the cluster and distance. They have provided the steps of the algorithm. Parallel kmean is developed to solve the problems as communication overhead and load balancing.

Aastha joshi et al [13] has reviewed that hierarchal clustering is the process in which given data sets is decomposited into hierarchal tree form. The decomposition of datasets into tree structure is done by two approaches. Top down and Bottom Up approach. Every cluster contains the child cluster and sibling cluster. Bottom up is more used as compared to the top down.

Sanjay et al [14] presents the original DBScan and the incremental approach of the DBScan. It is the density based notation of cluster. Density based spatial clustering of application with noise is used to extract clusters of arbitrary shapes in noisy large databases. By measuring the density of the point we identify the clusters. They presented the formula and algorithm for the incremental DBscan. Incremental approach result is better than the original one.

Ari et al [15] has presented the perch algorithm that is a non greedy algorithm. Mainly used for large scale clustering (online hierarchical clustering).This algorithm logically organize the data points to the leaves of an incrementally leave trees. They create

the cluster tree and uses rotation to correct the mistakes while encouraging the sallow tree. As result it concluded that perch algorithm is better than other clustering algorithm in terms of speed and accuracy.

Matioli et al [18] has presented the new algorithm for clustering that is based on univariate kernel density estimation, known as clusterKDE. This algorithm is used to cluster the datasets with the help of univariate KDE and optimization techniques. This algorithm is having the advantage of, do not require no of cluster as input argument. As per the result they described that clusterKDE is more strong and tough than cluster data algorithm and optimistic when it compared with the K-mean algorithm. They concluded that for clustering huge sample size data the clusterKDE algorithm is best. ClusterKDE is fast as compared to others.

Reference No:	Title of the paper	Methodology/ Algorithm	Advantages	Disadvantages	Work on map reduce
11	A Comparative Study of Various Clustering Algorithms in Data Mining	Kmean	Faster and better performance than other clustering algorithms	Very sensitive for noise large datasets	no
12	A Comparative study on Parallel Data Mining Algorithms using Hadoop Map Reduce: A Survey	Parallel k mean	There is no data movement and each processor compute distance independently	Communication overhead in exchanging the sum of squared errors	Yes
13	A Review: Comparative Study of Various Clustering Techniques in Data Mining	Hierarchal clustering	Tree of clusters are great for visualization and provide hierarchal nature between clusters	When one step split or merge is done it can never be undone, sensitive for noisy data	No
14	Analysis and Study of Incremental DBSCAN Clustering Algorithm	DBSCAN and its incremental approach	Suitable for large multi dimensional database. Time saving and effort efficiently.	Does not perform well on small databases. And insensitive in ordering of points	No
15	An Online Hierarchical Algorithm for Extreme Clustering	perch	Perch performance is best in terms of accuracy, quality and speed as compared to other algorithms.		no

IV CONCLUSION:

In this paper we have given the review of many algorithms, methods of classification and clustering algorithm of data mining. Some applications are applicable on map reduce also. We concluded that Parallel classification algorithms are more effective on map reduce as compared to the parallel clustering algorithms. We have discussed about the advantages and disadvantages of each algorithms as shown on the table format.

REFERENCES

1. Chen, Feng, et al. "Data mining for the internet of things: literature review and challenges." International Journal of Distributed Sensor Networks (2015).
2. Bin, Shen, Liu Yuan, and Wang Xiaoyi. "Research on data mining models for the internet of things." Image Analysis and Signal Processing (IASP), 2010 International Conference on. IEEE, 2010.

3. Shweta Bhatia Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 11, (Part - 1) November 2015, pp.82-85
4. Paul, Sujni. "Parallel and distributed data mining." New Fundamental Technologies in Data Mining. InTech, 2011.
5. Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 (1985): 580-585
6. Anyanwu, Matthew N., and Sajjan G. Shiva. "Comparative analysis of serial decision tree classification algorithms." International Journal of Computer Science and Security 3.3 (2009): 230-240.
7. Liu, Mingyang, Ming Qu, and Bin Zhao. "Research and Citation Analysis of Data Mining Technology Based on Bayes Algorithm." Mobile Networks and Applications 22.3 (2017): 418-426.
8. Srivastava, Anurag, et al. "Parallel formulations of decision-tree classification algorithms." High Performance Data Mining. Springer US, 1999. 237-261.
9. Wu, Gongqing, et al. "MReC4. 5: C4. 5 ensemble classification with MapReduce." ChinaGrid Annual Conference, 2009. ChinaGrid'09. Fourth. IEEE, 2009.
10. He, Qing, et al. "Parallel implementation of classification algorithms based on Map Reduce." Rough Set and Knowledge Technology (2010): 655-662.
11. Verma, Manish, et al. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2.3 (2012): 1379-1384.
12. Lokeswari, Y. V., and Shomona Gracia Jacob. "A Comparative study on Parallel Data Mining Algorithms using Hadoop Map Reduce: A Survey." Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2016.
13. Joshi, Aastha, and Rajneet Kaur. "A review: Comparative study of various clustering techniques in data mining." International Journal of Advanced Research in Computer Science and Software Engineering 3.3 (2013).
14. Chakraborty, Sanjay, and Naresh Kumar Nagwani. "Analysis and study of Incremental DBSCAN cluster algorithm." arXiv preprint arXiv: 1406.4754 (2014).
15. Kobren, Ari, et al. "An Online Hierarchical Algorithm for Extreme Clustering." *arXiv preprint arXiv:1704.01858* (2017).
16. Zhu, Fubao, et al. "A Classification Algorithm of CART Decision Tree based on MapReduce Attribute Weights." *International Journal of Performability Engineering* 14.1 (2018): 17.
17. Batra, Mridula, and Rashmi Agrawal. "Comparative Analysis of Decision Tree Algorithms." *Nature Inspired Computing*. Springer, Singapore, 2018. 31-36.
18. Matioli, L. C., et al. "A new algorithm for clustering based on kernel density estimation." *Journal of Applied Statistics* 45.2 (2018): 347-366.