International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 1
ISSN: 2454-4248
265 – 270

Analytics of IoT Streaming Data using Modified New Pattern Mining Algorithm

Monika Saxena[1], C. K. Jha[2]

muskan.saxena@gmail.com, ckjha1@gmail.com

Abstract. In the era of information technology, everything we are using in the everyday life is represented in form of information. Transportation, parking, traffic, pollution are some examples of hundreds of infrastructure systems with which we act every day. By using information technologies combined with communication, it becomes very easy to represent all details even the tiniest parts of these fields in forms of data. Furthermore, the Internet of things (IoT) plays a very important role in connecting physical objects with electronics, software, and sensors. Based on that, smart cites have been modeled and implemented in thousands place over all the world; In these cities, all smart systems in different fields like transportation networks, pollution,traffic,airlines, etc. are showed in form of numbers and strings of characters.This paper represents the problems occur in this type of methods with little bit solution of them by new modified algorithm.

Keywords: *IoT Streaming, Map Reduce, parallel data mining.*

*****

1. Introduction

In the Smart cities, all smart systems in different fields like transportation networks, pollution,traffic,airlines, etc. are showed in form of numbers and strings of characters. This data is collected to form a big data source. This data can be accessed from everywhere via internet to know some information or take a decision.Also, transport energy, health care and waste management are a good example for systems which were improved intelligently, smartly, and automatically [1][2]. The live Data is very important to know facts like knowing the status of things to take immediate decisions in various aspects of our daily life. Furthermore, it is very essential to store the historical data of these things for several past years. Thereby, new rules can be predicted; unknown behaviors can be deduced thanks to the aid of analytical algorithms. Having all these sources of big data such as emails, social sites, videos, images, blogs, sensor data which are produced daily from the infrastructural systems of various aspects of life such as transportations network, pollution, IoT, Traffic systems either for buses network, underground metrosnetwork, trains network, these massive amounts of data need very huge space of storage and very special parallel computing systems. The most important questionsarise at this point are that, how these immense amounts of big data can be stored and be processed? And how to fetch the meaningful data from millions of millions of records of data?In order to answer this question, it should be known that, not all pieces of the big data are important; a lot of them are redundant information. Consequently, filtering the data is the primary clue for solving this problem. By distinguishing the unique information and filtering the meaningful data, it could be save storage and processing time. On the other hand, by determining the frequent patterns of data, it could helpgreatly to predict the associate rule sets that can be taken as a guide in deducing the behavior of systems in advance based on the historical data. This approach is called data and frequent pattern mining.Distributed pattern mining is one of the solution method to improvethe performance of processing the Big Data. Furthermore, it saves exabytes of storage space alongside withsaving the processing time. Not only that but also, it widely opens the door for mining thousands of rule sets that are used in predicting facts and reveal the mysterious behavior of the unknown systems[3]. Data mining processis deployed by running some parallel programming tools like SAMOA (Scalable Advanced Massive Online Analysis) orMapReduce [4]. In the next section, the frequent patterns and data mining process and its evolutions is demonstrated including the good features and the weak points to be used as a guide in approaching a new technique or method of distributed pattern mining that keep on the advantages and avoid the drawbacks of the conventional techniques.

Internet of things is a machine to machine connectivity. 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].

We have implemented parallel data mining algorithm in order to improve the speed, accuracy and quality of the algorithm which is going to be applied for data mining from ware houses.

1.2 Pattern mining and Data Mining

Pattern mining algorithms are used to mine the useful patterns from the massive amount of datasets. Mostly used pattern 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.

Data mining is the techniques developed to handle wide amount of data using different tools to process it properly. It involves finding interesting and useful pattern from large data sets and to extract the hidden and useful information. The main reason behind the development of

 ISSN: 2454-4248
265 – 270

algorithm is to fetch meaningful information from wide amount of data. Data preparation is the method in which data is integrate from various sources and then prepared for mining. After cleaning algorithm will apply to evaluate patterns, at last the knowledge is represented to the user [1].
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].

1.2 Parallel and Distributed Data Mining

Sujni Paul [7] 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: This algorithm is used 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: In this algorithm data is distributed in the different 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".

2 Problem Formulation & Proposed System

According to Literature survey, we require a lot of solutions or updation for variety, volume and velocity of data. Here, we state basic problems that researchers are going to overcome using new algorithm.

- Mining of Dynamic Big Data
- Scheduling of Dynamic Big distributed data for high performance.
- Discover hidden pattern in complete data set that is partitioned and physically distributed

In Analyzing and process Dynamic Big data, we have challenges in terms of scheduling, decision making and High performance. If we focus on variety and volume of dynamic big data, then we have one solution that is distributed data mining.

2.1 Conventional Algorithms

Today"s, internet of things brings the grand challenge for analysis, management and proper extraction of knowledge from the data houses. To maintain the data quality for providing, the high best quality and faithful data so that the best decision would be taken for business level analysis. There are many algorithms were used for good quality data, to remove inaccurate data and noises present in the information data [2] as shown in figure

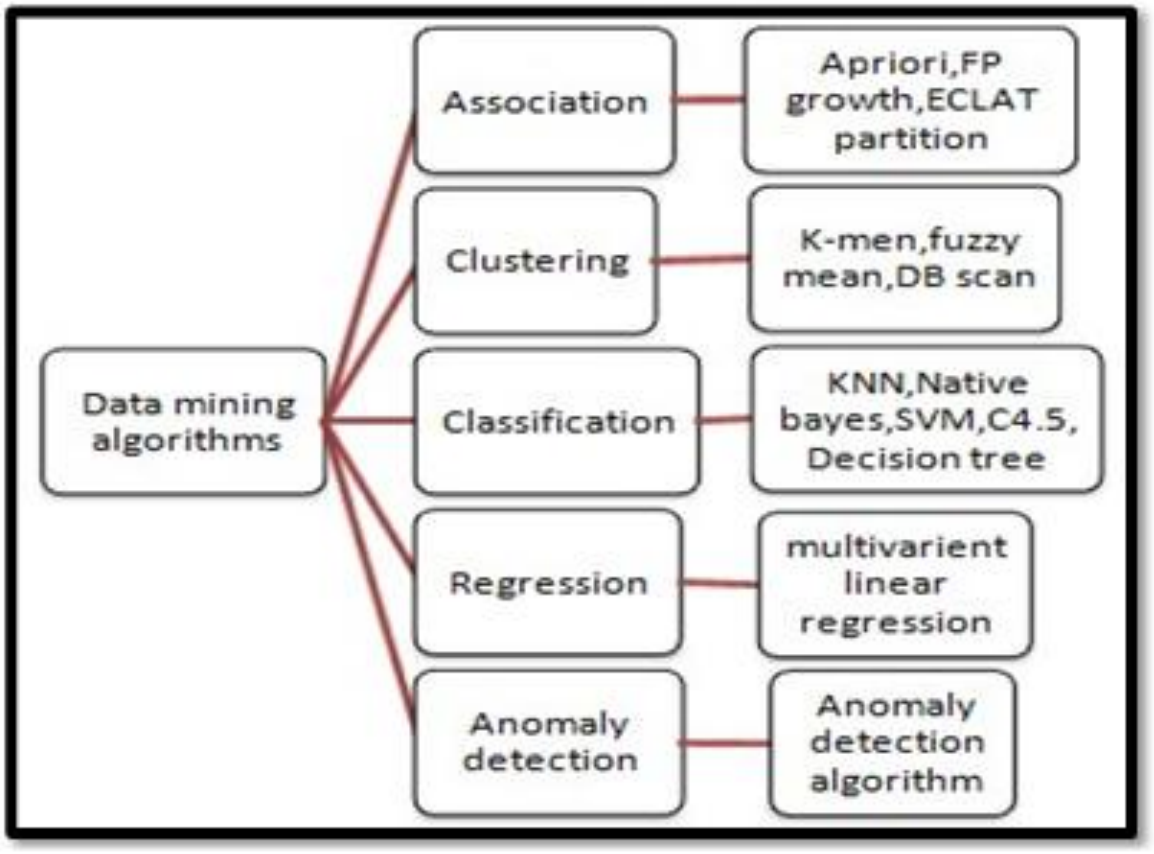

Figure 1: Classification of different Data mining algorithms

 ISSN: 2454-4248
265 – 270

Association rule displaying attribute value conditions that frequently occur together in the given set of data. It focuses on „market basket analysis" [3].To generate association rule Apriori algorithm is helpful. Apriori assumes that only items within transaction or item sets were sorted in lexicographic order. The Apriori algorithm generally performs join and prune step and then calculate the frequency of those candidate who were generated by scanning the database [3].

Data clustering algorithm divides the data into meaningful groups so that the patterns in the same group must be similar with some sense and vice versa. Clustering methods are divided into four parts: portioning, hierarchical, density and grid based method. The advantage of this method is its fast processing time that is independent of the no of data points and depends only on no of cells [2].

Data classification is the management of decision making, we have an object, assigning it to one of predefine target category is classification [3]. 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

Regression is used to determine the relation between the dependent variable and independent variables the independent variables may be more than one [10]. The process in which it identifies the relationship and the effects of this relationship on the future outcomes value of data is defined as Regression. Mostly used regression Algorithms are Simple linear regression, Lasso regression, Logistic regression, Support vector machine, Multi variant linear regression.

Anomaly detection use the anomaly detection algorithm to remove the unwanted noise and in complete data from the database. It is used to find suspicious cases that are not usual within the data. It is an essential tool for fraud detection, network intrusion and rare events that are hard to find but have more significant. It is also the form of classification.

We are going to work on two algorithms of association rule for finding the frequent pattern we will apply these two algorithms on smart city datasets and will compare new novel algorithm with them:

1. Apriori Algorithm
2. FP Growth Algorithm

2.2. Distributed frequent patterns and Data mining

Data volume that is represented in size of data, Data Variety which is expressed in terms of type of data and its homogeneity, velocity that is represented in the form of data, and the value of Data which is measured in how much the meaningful of data, are the great challenges in performing any type of analysis in real time. On top of these, is the veracity of the data. Frequent pattern mining is to determine all relationships in between the items of data in a dataset. For example, assuming there is adatabase ADB with transactions Tid1 ...TidN, calculating all patterns „P" that are present in at least a fraction „r" of the transactions when"r" refers to the minimum support which is the less number of repetitions of a pattern of itemset. The parameter „r" can be expressed as an absolute number and it can be defined as a fraction percentage equivalent of the number of transactions in the ADB. All transaction Tidk considered as a vector of binary value and collection of discrete values. It representing the identifiers of the attributes (binary) that contains start value 1.

Consider the set of items are M and D. The rule M ⇒ D is considered as a rule at minimum support „r" and minimum confidence „c", when the following two conditions are true:

I. The set M ∪ D is a frequent pattern.

II. The ratio of the support of M ∪ D to that of M is at least „c".

The minimum confidence „c" is always less than 1 because the support of the set M ∪ D is always less than of M [5].

The main idea behind deciding the frequent patterns, consists of three main sub tasks which are: scanning the database for all distinct items value in the database to build the list of candidates, making all possible itemset patterns of these candidates to build the complete set of all possible itemset patterns, and testing the frequent of each pattern throughout the whole database transactions. There are many algorithms for calculating the frequent pattern. For comparing the performance parameters of these algorithms, there are two main parameters which are the number of iterations or looping paths through the database transactions and the number of candidate sets. the less are the number of looping paths and the candidate set.

The more better is the performance.These algorithms will be surveyed later in the literature review section in order to show the proses and cones in order to keep the good features and avoid the drawbacks rather than adding an improvement in the proposed technique.

The most two common algorithms for determining the frequent patterns are the Apriori algorithm which was proposed by Agrawal and Srikantin 1994 to fix the problem of mining frequent itemset[6]. And, The FP-growth method which was developed by Han et al., [7] which utilizes the FP-tree data structure to store the frequency of information in the transaction database. FPgrowth applies a frequent divide-and-conquer method and the database projection approach to determine efficiently the frequent itemset without candidate generation like in the Apriori algorithm,.

The most of frequent pattern set mining algorithms are eitherApriori-like algorithm or FP-growth-like algorithm. All of them have been used in the field of

 ISSN: 2454-4248
265 – 270

analysis of patterns. Researchers found that these algorithms which are like the Apriori algorithm,have some limitations such as candidate key generation and number of times repeated scans of the whole transactions of the database, which requires candidate tests too. Therefore, in case of there is very large and complex datasets, the time and complexity are increased proportionally. While The FP-growth like algorithms appliesthe „Divide and Conquer" strategy and does not require candidate key generation tests. Furthermore, it does needs only two paths of scan of the database transactions. therefore, it can be concluded that the FP-growth algorithm has a faster performance.

However, the FP-Growth algorithm requires a much bigger space of storage to store the information of the nodes of the tree structure and the liking pointers and indexes. Consequently, there is a room for enhancing either the performance of determining the frequent pattern or minimizing the space of storage that is required for the mining process.In this research we are proposing a new distributed pattern mining algorithm to overcome the performance or the space of storage related problems in parameters like minimizing the number of candidates set andminimizing the number of paths of scanning the database transactions which in turn reducing thecommunication overheads,throughput, memory usage & computational costs of I/O of Big Data Mining. The new method will be categorized as Apriori-like method. However, it is completely difference in searching techniques and forming the candidates and itemset lists. The proposed method can complete the process efficiently in just one path of scan of the database transactions with a significant reduction in the number of frequent sets and in turn reducing the number of tests for the frequent of the pattern set. Moreover, the only one path of scan, applies the binary search technique which has a complexity of N LOG N while the traditional Aproiri uses many paths of scan applying the linear search algorithm with complexity of N^2
. Furthermore, it will save the required space for a tree structure and its nodes and linking pointers which is the drawback of the FP-Growth method.

2.3 Evaluation method

In this research we are proposing a new distributed pattern mining algorithm to overcome the performance or the space of storage related problems in parameters like minimizing the number of candidates set and minimizing the number of paths of scanning the database transactions which in turn reducing the communication overheads, throughput, memory usage & computational costs of I/O of Big Data Mining. The new method will be categorized as Apriori-like method. However, it is completely difference in searching techniques and forming the candidates and itemset lists. The proposed method can complete the process efficiently in just one path of scan of the database transactions with a significant reduction in the number of frequent sets and in turn reducing the number of tests for the frequent of the pattern set. Moreover, the only one path of scan, applies the binary search technique which has a complexity of N LOG N while the traditional Aproiri uses many paths of scan applying the linear search algorithm with complexity of N^2 . Furthermore, it will save the required space for a tree structure and its nodes and linking pointers which is the drawback of the FP-Growth method.

3. Simulation and Results

Proposed solution is simulated with new modified algorithm, Our test deployment of now able to support

1. Input/output Load
2. Traffic Control

3.1 Test case: The IoT Big Data test:

In smart cities, hundreds of thousands of sensors are working all the time to log the physical measurements of humidity , ambient temperature , pressure , wind direction , wind speed , viscosity .. etc. these pieces of information have been logged every specific amount of time say every 10 minutes or whatever. These massive amount of data have been collected from the below data source:
"*http://iot.ee.surrey.ac.uk/citypulse/datasets/weather/aarhus_weather_dewpoint"*

3.1.2 Data samples

The raw data of IOT having fields like measurements of the temperature, humidity, viscosity, wind speed, and wind direction.

3.1.3. Data preparation

In order to have the data prepared to be processed by any frequent pattern algorithms it needs to be as follow:

1- All fields are formatted in numeric format

2- Every column has a unique numeric range

The encoding procedure will be as follow:

Converting the timestamp column to numeric format

Keeping the humidity column as itis

keeping the dwpoint column as it is

keeping the pressure column as it is

adding 2000 to the temperature column

adding 4000 to the winddirection column

applying this formula for the windspeed=windspeed*10+6000

by applying this procedure to the raw data, the prepared data will be ready for processing by frequent pattern algorithms. Table 5.1 shows a 60s record as a sample records of the prepared data of the IoT big data.

3.1.4 Test Results

 ISSN: 2454-4248
265 – 270

Dataset Name **IOT sensors data**
Transaction Count **121844 TIDs**
Time of processing (ms)

Support(%)	Apriori	fpgrowth	New algorithm	FP itemsets count	Minimum support in transaction
5	859	782	1063	15	6092
1	4688	1312	5822	214	1218
0.5	6422	1093	5585	519	610
0.25	17965	1078	5948	1432	305
0.1	17816	1172	6091	3418	122
0.05	28701	1187	6982	8182	61

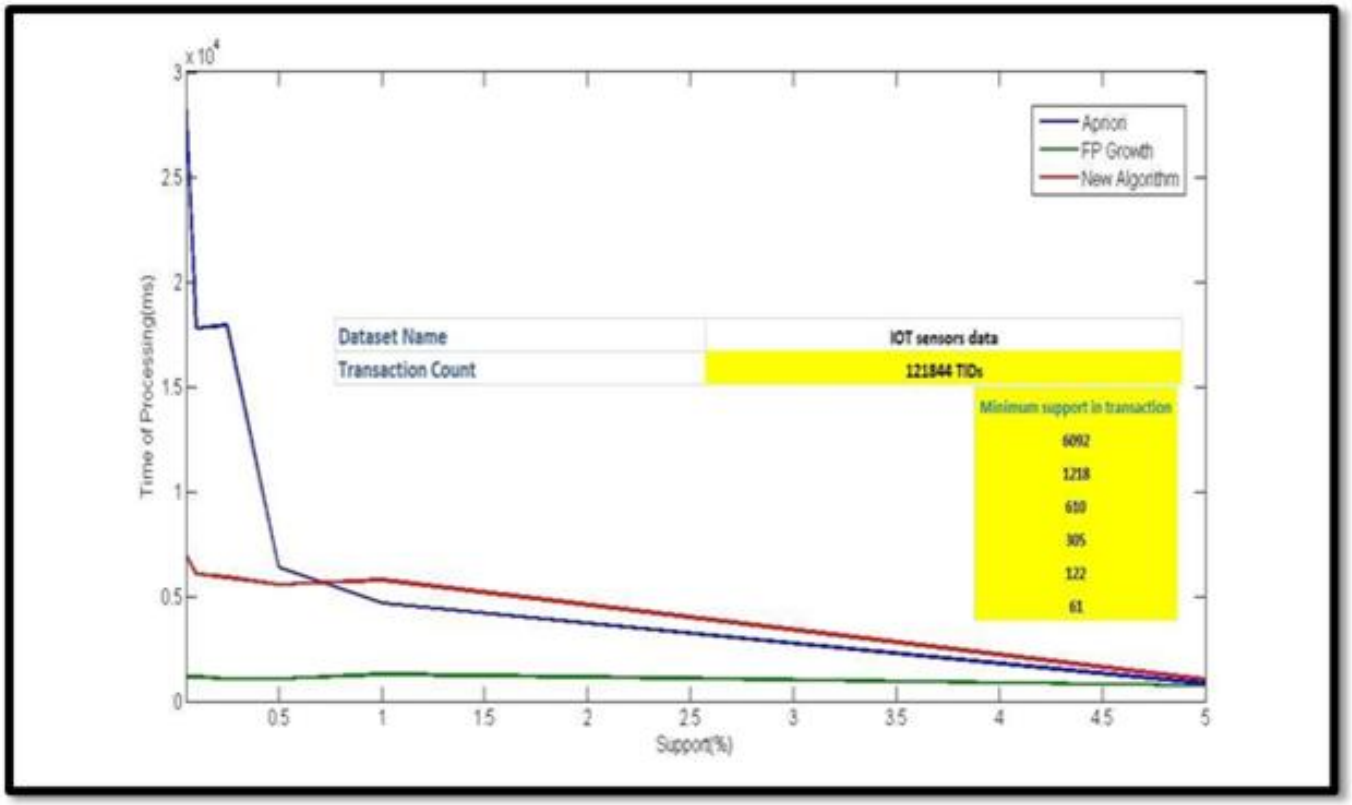

Figure 1 Input/output Load after using algorithms.

 ISSN: 2454-4248
265 – 270

6. Conclusion and Future work

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] hen, 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).