International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 07 | July 2019 www.irjet.net p-ISSN: 2395-0072

Improving the Performance of Smart Heterogeneous Big Data

Divyanshi Mahajan[1], Parveen Singh[2], Vibhakar Mansotra[3]

[1]Student, Dept. of Computer Science and IT, Jammu University, Jammu and Kashmir, India
[2]Associate Professor, Govt SPMR College of Commerce, Jammu and Kashmir, India
[3]Professor, Dept. of Computer Science and IT, Jammu University, Jammu and Kashmir, India (Supervisor)

Abstract -*Data mining is a popular and promising field and its growing impacts can be visualized in almost every aspect of life. Data mining is an interdisciplinary subfield of computer science that discovers the hidden patterns in the large data and depending on the degree of coupling between data and data mining techniques, different types of Data MiningSystems are defined. The rise of big data is revolutionizing the economy, as it has commenced learning complex models with millions to billions of parameters for gaining valuable insights that are transforming businesses and in many aspects of our lives like smart city, smart home, Industries, etc. using Internet Of Things. Big Data containing an enormous amount of data is processed, stored and analyzed using association rule mining algorithm which is Apriori, FP-Growth. One of the challenges is to understand the meaningful data called as Smart Data, which we get using association rules the best of the rules are generated for getting the knowledge of Smart Big Data. The results for Apriori Model and the FP-Growth model are scrutinized by the performance of big data in terms of Execution time, Processing Time, Memory for different Support count and same confidence. For improving the performance of Big Data we have Proposed a method in which we have used one of the feature selection method Principal Component Analysis and analyze it for both Apriori and FP-Growth. The data is processed using the RapidMiner tool and the result is scrutinized based on Execution time. It is examined that the proposed algorithm gives good performance in comparison with the earlier algorithm.*

Key Words: **Big Data, RapidMiner, Apriori-algorithm, FP-Growth algorithm, PCA .**

1. INTRODUCTION

Big Data is the most favorable topic these days. When the data in high speed with huge quantity is present in Gigabytes, Terabytes, etc is to be processed and analyze through various sources, this type of data is known as Big Data.

Big data mining is the data mining technique that is performed on the large volume of data. Apart from data collection it also provides the way to the gathering, storing, organizing and analyzing data collected from different sources. Volume, Variety, Velocity and Veracity are the most common dimension of Big Data. A large volume of data (volume) coming from multiple sources can enter the cloud under different formats (variety) and can demand to process in real-time (velocity) with high levels of accuracy (veracity) [1]. The challenge for Big Data is to understand meaningful data from the complex large data refer as "Smart Data". Therefore, Smart data is getting meaningful and filtered data from Big Data.

Heterogeneity is an important aspect of big data. Heterogeneity in terms of big data means to deal with structured, semi-structured, and unstructured data. Heterogeneous data contains different data types and formats within the data. The data present is of poor quality due to missing values and high data redundancy. Structured data is stored in rows and columns with specific schemas. Following applications like Customer Relationship Management and Enterprise Resource Planning systems create structured data. Semi-structured data consist of metadata that describes the structure of data not always fit in rows and columns, the data is produced using the sensor, web feed, event monitor, stock market feeds and security system. In Unstructured data, the Data is produced in a large amount of volume from different sources at a high speed with no accuracy. This type of data is present in various forms such as social media, text documents, videos, audio, and images.

Big data is a huge amount of data for mining is present in data ware houses and databases. In Big data, data mining can be used for extracting or recognizing the patterns and getting information from the data. Data mining techniques are classification, clustering, association rules, prediction, estimation, documentation and description which might be functional to big data. The investigation of these techniques was done long ago. For big data from all data mining techniques the association rule mining is the most efficient data mining technique. This data mining technique is used to determine a range of unknown patterns and knowledge from big data. Hence, the analysis and for finding the correlation between various aspects of data is done using the association rule mining algorithm.

2. Associative Rule Mining

The mining of association rule is done to find correlation, association and frequent patterns between different set of items in database and information regarding data repositories. In Association Rule Mining, one item set correlates with another set of items in the same transaction. Figure 1 depicts frequent item set and their following subset. Association Rule is measured using support and confidence

values and generating best association rules, as defined below.

- *Support* defined as how frequently the set of items occur together in data.
- *Confidence* defined as how frequently the rule is found to be true. We can say for considering useful associations between the items.

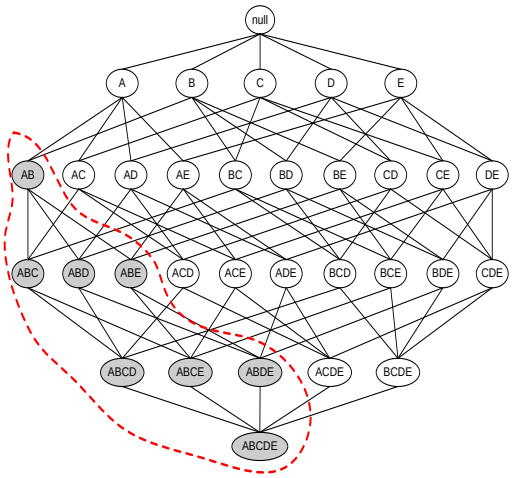


Figure 1: Frequent itemset and its subset

Let N be the number of records in a database, X(I) be the number of records with item set I, M be an item set with g elements $M_1, M_2, \ldots, M_g$ and N be an item set with h elements $N_1, N_2, \ldots, N_h$. An Association Rule M⇒N can be generated if the support of M and that of N is above the minimum support value and also the confidence of the rule M⇒N is above the minimum confidence specified. Support S, of M is a probability that a transaction contains M and is given in (1) and (2).

$$S(M) = P(M) \quad (1)$$

$$S(M) = X(M) / X \quad (2)$$

Support of M⇒N is a probability that a transaction contains both M and N as given in (3) and (4).

$$S(M \Rightarrow N) = P(M \cup N) \quad (3)$$

$$S(M \Rightarrow N) = X(M \cup N) / X \quad (4)$$

Confidence C, of M⇒N is a conditional probability that a transaction that contains M contains N also. Confidence can be calculated as in equations (5), (6) and (7).

$$C(M \Rightarrow N) = P(M \mid N) \quad (5)$$

$$C(M \Rightarrow N) = X(M \cup N) / X(M) \quad (6)$$

$$C(M \Rightarrow N) = S(M \cup N) / S(M) \quad (7)$$

Association Rule Mining consists of two steps to find frequent item sets and generating association rules. The set containing all item sets are called as candidate item sets. The item set which occurs frequently in any data are known as frequent item sets. The frequent item sets satisfy the minimum support(min_sup) that is user specified . If the min_sup is 5%, then the item sets whose support % are greater than or equal to 5 are considered as frequent item sets. The association rules are generated using the frequent item sets and they should satisfy the minimum confidence (min_conf) of 50%.

2.1 Apriori Algorithm

Apriori algorithm is used to find frequent itemsets in data and to generate Association Rules from the frequent itemsets. Apriori algorithm uses level-wise search to find frequent itemsets. In Apriori algorithm, x-itemsets are used to explore (x+1) itemsets. An itemset containing x items is known as x-itemset. In this algorithm, frequent subsets are increased one item at a time, this step is known as the candidate generation process and then group of candidates are evaluated towards the data. For counting candidate itemsets accurately, Apriori uses breadth-first search method and a hash tree data structure. Atlast identification of the frequent individual items in the database is done and continues them to larger and larger itemsets as long as those itemsets appear.

2.2 FP-Growth Algorithm

FP-Growth algorithm uses divide-and-conquer for mining frequent itemsets without candidate generation. The first step, it reduces the database describing frequent items into a frequent pattern tree, known as FP-tree, which helps in retaining the itemset association information. Next step is to divide the reduced database into a set of conditional databases, each linked with one frequent item or pattern and obtaining each conditional database separately. Over each pattern fragment, only its associated data sets need to be examined. Therefore, this approach reduces the size of the itemsets to be searched, simultaneously.

3. RELATED WORK

Big Data involves complex growing large volume datasets with multiple independent sources. With the fast growth of networking, storage of data and collection capacity, Big Data is developing rapidly in all science and engineering domains. Xindong Wu et al[2] depicts the theorem HACE (Heterogeneous, Autonomous, Complex and Evolving) that defines the characteristic of Big data revolution, and proposed a Big Data processing model from the data mining prospect. Overall, it examines the challenging effects of the data-driven model. But it needs the testing and performance evaluation that relates to any dataset.

Big Data will continue growing over the upcoming years, and data scientists have to manage enormous amounts of data every year[3]. The data is going to become more diverse, larger, and faster. Some insights about the big data, and the

main concerns, and the main challenges for the future. Big Data is becoming the new Final Verge for scientific data research and for business applications.

Data mining technologies and their comparative study of the Internet of Things, which comprises of clustering, classification and pattern mining technologies, from the perspective of infrastructures and their services are defined[4]. This paper shows the challenge to collect heterogeneous data.

For better serving the demands of web-based applications, Web usage mining is the application of data mining techniques for identifying useful patterns from web data. K.R.Suneetha , et. al [5] displays the depth analysis of Weblog data analysis of the NASA website. This encouraged the researchers to explore Web usage mining using Big Data.

In present days many industries and governments are using Big Data to extract valuable perceptions. Such perceptions can help decision-makers to enhance their strategies and upgrade their plans. It helps the organization to provide added value for many economic and social sectors and also to gain competitive benefit. Benjelloun et al [6] exhibited many Big Data projects, opportunities, examples, and models in many sectors such as health care, commerce, tourism, and politics. It concludes that the Big Data revolution contributes to enhancing different scientific fields by allowing advance complex analysis across multiple sources.

Apriori, the first Association Rule Mining algorithm, was proposed by Agrawal[7], and it strongly diminished the search space size with a downwards closure Apriori property that says a n-itemset is frequent only if all of its subsets are frequent. This is described by a feature called candidate generation, where (n + 1) candidate itemsets are generated repeatedly by merging any two frequent n-itemsets which share a common prefix of length (n-1). Moreover, the calculation of support of each candidate itemset is then performed to determine if the candidate is frequent or not. Finally, the algorithm stops if frequent itemsets are not generated.

Association rule is a well-known mining technique in data mining. It is the inference of correlation of events or objects and their correlation can be represented as rules for the ease of understanding and the convenience for applying the rules to predict the occurrence of an event or object in the future. There are many efficient techniques for performing the mining using association rule, such as Apriori [7], Eclat [8], and FP-growth[9].

Frequent Itemset Mining was conferred by R. Agrawal and R. Srikant in 1995 [11]. Frequent Itemset Mining is the extension of their research work done in 1994[10]. The paper introduced two Apriori-based algorithms called Apriori-All and Apriori-Some. Apriori algorithms are adapted with Map Reduce by doing multiple iterations of Map Reduce job and generating candidate itemset and finding them in the database. But the process of scanning Database and multiple Map Reduce job execution is very expensive. This limitation becomes more severe when dealing with large data sets.

4. METHOD AND MODEL

The research work has been done by using tool RapidMiner studio 9.3.001 and the following work is done as describe below.

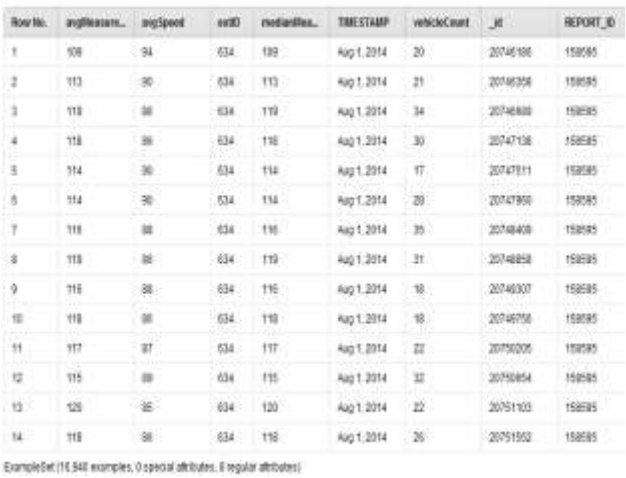
4.1 Dataset

The dataset is mainly based on secondary data collected from Citypulse Smart city datasets of Aarhus city in Denmark. This benchmark data is retrieved from http://iot.ee.surrey.ac.uk:8080/datasets.html. The analysis carried out in this paper is based on the 2 months data of traffic to get the best and efficient rules.

In smart cities, Traffic can be monitored by using variety of sensors that logs hundreds of physical measurements to count the number of cars and their speed in each road at any timestamp. A collection of datasets of vehicle traffic, observed between two points for a set duration of time over a period of 2 months. In this dataset , the number of cars are counted by sensing the Average speed in time , Average Measured Time, by allotting different id to every vehicle and by Timestamp of 10 minutes for all roads .

4.1.1 DATA COLLECTION

The collection of traffic dataset from Citypulse website and of the region Aarhus city in Denmark. The total number of records present is 16,940 for the traffic dataset and sample of the raw traffic dataset is shown in figure 2.



Row No.	avgMeasure...	avgSpeed	extID	medianMea...	TIMESTAMP	vehicleCount	_id	REPORT_ID
1	108	94	634	109	Aug 1, 2014	20	20746186	158585
2	113	90	634	113	Aug 1, 2014	21	20746358	158585
3	118	88	634	119	Aug 1, 2014	34	20746888	158585
4	118	89	634	116	Aug 1, 2014	30	20747136	158585
5	114	90	634	114	Aug 1, 2014	17	20747311	158585
6	114	90	634	114	Aug 1, 2014	29	20747950	158585
7	116	88	634	116	Aug 1, 2014	35	20748409	158585
8	119	86	634	119	Aug 1, 2014	31	20748858	158585
9	116	88	634	116	Aug 1, 2014	18	20749307	158585
10	119	86	634	119	Aug 1, 2014	18	20749756	158585
11	117	87	634	117	Aug 1, 2014	22	20750205	158585
12	115	89	634	115	Aug 1, 2014	32	20750654	158585
13	120	85	634	120	Aug 1, 2014	22	20751103	158585
14	118	86	634	118	Aug 1, 2014	26	20751552	158585

ExampleSet (16,848 examples, 0 special attributes, 8 regular attributes)

Figure 2: Sample of the raw data of traffic dataset

4.1.2 DATA PREPROCESSING

Data is processed by removing of the unwanted attributes and taking only useful attributes . For traffic dataset the following attributes to be used for getting good results are

AvgMeasuredTime, AvgSpeed ,Timestamp, Vehiclecount and id as shown in figure 3.

Row No.	avgMeasure...	avgSpeed	TIMESTAMP	vehicleCount	_id
1	109	94	Aug 1, 2014	20	20746185
2	113	93	Aug 1, 2014	21	20746358
3	119	86	Aug 1, 2014	34	20746689
4	118	88	Aug 1, 2014	30	20747138
5	114	90	Aug 1, 2014	17	20747511
6	114	90	Aug 1, 2014	28	20747960
7	116	88	Aug 1, 2014	35	20748409
8	119	86	Aug 1, 2014	31	20748859
9	116	88	Aug 1, 2014	18	20749307
10	118	86	Aug 1, 2014	18	20749758
11	117	87	Aug 1, 2014	22	20750205
12	115	89	Aug 1, 2014	32	20750654
13	120	85	Aug 1, 2014	22	20751103
14	118	86	Aug 1, 2014	25	20751552

ExampleSet (16,940 examples, 0 special attributes, 5 regular attributes)

Figure3:Sample of resultant traffic dataset

4.2 Proposed Model (Improvement Using PCA)

For improving the performance of the smart heterogeneous big data Principal Component Analysis (PCA) is used .As the data containing interrelated features are reduced by PCA. For transforming the dependent features to a new set of independent features .The term Principal component is a smaller number of features collected from the most appropriate information . Each component is combined from a linear function of the variance-covariance matrix of original dependent features. The analysis provides the the correlations between each principal component and the original features percentage of variance explained by each Principal component [13].The two proposed models are

a. Using PCA and Apriori
b. Using PCA and FP-Growth

4.3 EXPERIMENTATION

The proposed research is implemented in RAPIDMINER and the results are evaluated by comparing proposed and existing methods with respect to certain performance measures.

4.3.1 Test case1:

W-Apriori process is an extension of Weka-Apriori in RapidMiner. In the following process shown in figure4.1 the data(in csv) is retrieved from the local repository in this repository the data is saved by importing it from our computer and then drag it from the local repository and drop it in process window then retrieve() process is formed[10]. Next, discretize by frequency operator is used for data transformation that is it converts numerical attributes to nominal attributes. Then date to numerical operator is used for converting Timestamp attribute of traffic dataset into numerical attribute. As W-Apriori works only for non-numerical attributes so we have used Numerical to Binomial operator and lastly Log operator, it stores information into

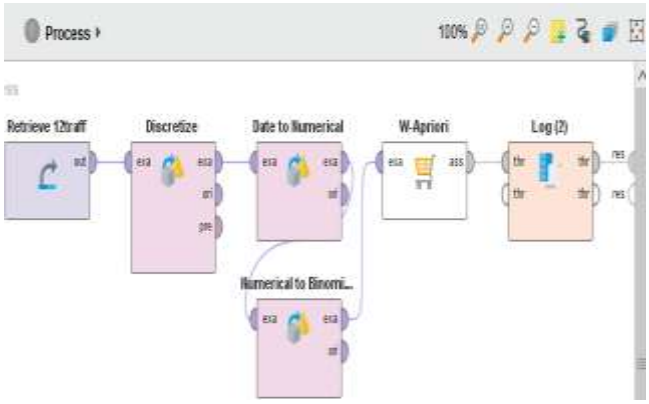


Figure 4.1:Process for generating rules using W-apriori

the table. This information can be almost anything including parameter values of operators, memory, execution time etc. Then this log operator is connected to the result port for getting results.

4.3.2 Test Case 2:

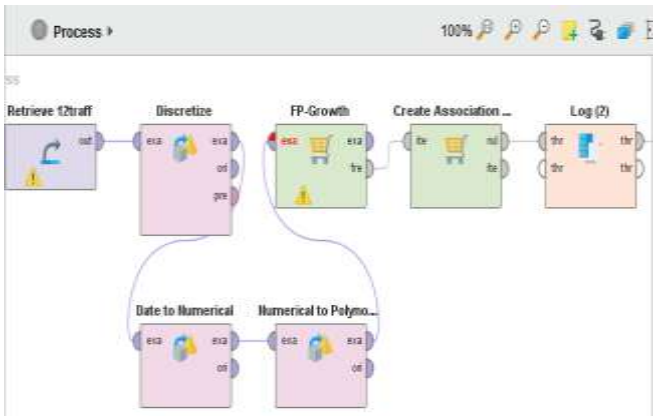


Figure 4.2: Process for generating rules using FP-Growth algorithm.

The following process as shown in figure 4.2 is used for generating rules and performance measures. The process of FP-Growth for the following dataset is as described. Firstly the data(in csv) is retrieved by dragging it from the local repository and drop it in process window then the retrieve() process is formed. Then, discretize by frequency operator is used for data transformation that is it converts numerical attributes to nominal attributes. Then date to numerical operator is used for converting Timestamp attribute of traffic dataset into numerical attribute. A FP Growth works for non-nominal attributes so we have used Numerical to Polynomial operator. After FP Growth we have used 'Create Association Rule' operator for generating association rules as shown in figure4.3. Then Log operator is used for storing information into the table. This information can be anything including parameter values of operators, memory, execution

time etc. Then this log operator is connected to the result port for getting results.

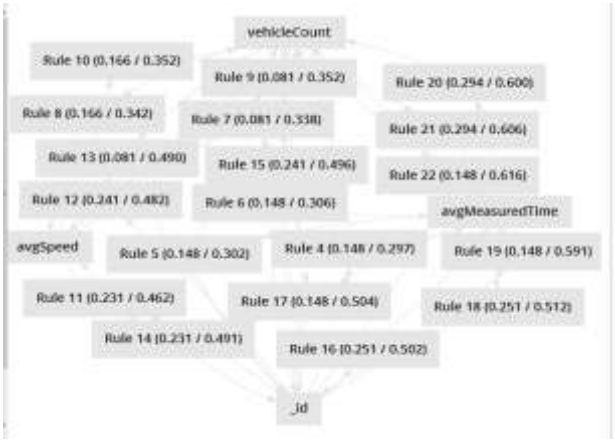


Figure4.3:Graph for the creation of association rules

4.3.3 Test Case 3:

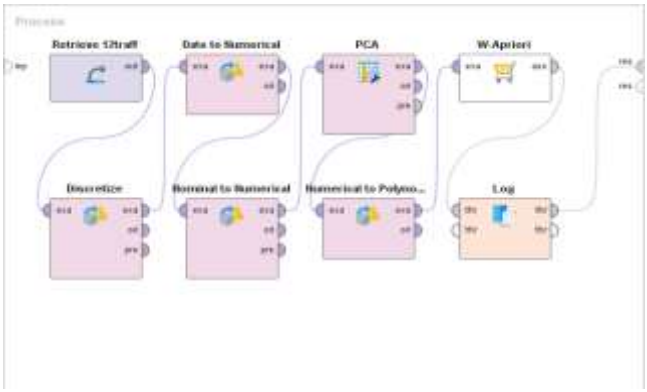


Figure4.4 :Process for improving performance using PCA+W-Apriori .

For improving the performance of our datset using W-Apriori we have used one of the best feature selection method PCA (Principal Component Analysis) that is using PCA with W-Apriori, the proposed model is shown in figure4.4 for PCA and W-Apriori.

4.3.4 Test Case 4:

For improving the performance of our dataset using FP-Growth we have used one of the best feature selection method PCA (Principal Component Analysis) that is using PCA with FP-Growth, the proposed model is shown in figure4.5 for PCA and FP-Growth.

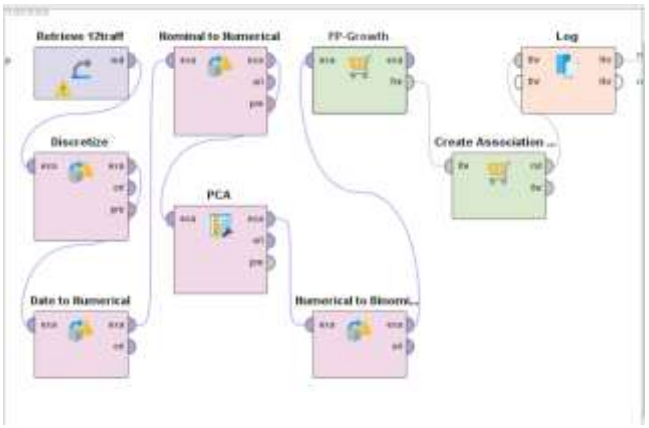


Figure 4.5: Process for improving performance using PCA+FP-Growth

5. RESULTS

All the results have been evaluated with different support count and same confidence. Based on parameters Execution time, Processing time and Memory as shown in Table 1.1 For Support 5% , Table 1.2 For Support 1% , Table 1.3 For Support 0.5%.

Table -1: Support 5%

Parameters	Execution time(ms)	Process time(ms)	Memory (mb)	Confidence
Apriori	140	453	1219.81	0.2
FP-Growth	452	562	701.29	0.2
PCA+Apriori	78	562	810.28	0.2
PCA+FP-Growth	140	440	835.63	0.2

Table -2: Support 1%

Parameters	Execution time(ms)	Process time(ms)	Memory (mb)	Confidence
Apriori	110	281	907.74	0.2
FP-Growth	281	390	502.38	0.2
PCA+Apriori	62	281	1148.72	0.2
PCA+FP-Growth	93	249	1200.21	0.2

Table -3: Support 0.5%

z	Execution time(ms)	Process time(ms)	Memory (mb)	Confidence
Apriori	78	187	1068.31	0.2
FP-Growth	172	281	748.16	0.2
PCA+Apriori	78	249	1029.37	0.2
PCA+FP-Growth	94	250	866.70	0.2

Execution time is more reliable than memory and processing time. Therefore for improving the execution time of apriori and fp growth , PCA along with apriori and fp growth is used .The results are shown in the Chart 1.

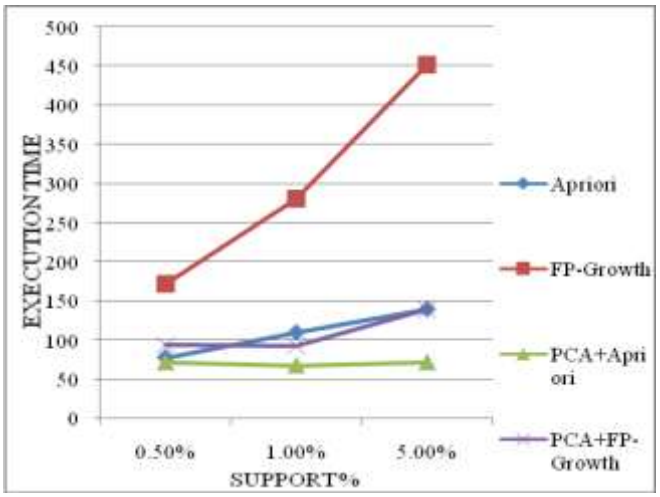


Chart -1: Execution time comparison

From the figure the execution time of PCA and Apriori and PCA and FP-Growth has improved as compared to the execution time of Apriori and FP-Growth. Therefore with the decrease in support count, the execution time also decreases.

6. CONCLUSION

Association rule mining is efficiently used for high dimensional datasets, The main aspect for both frequent pattern algorithm is its Execution time. For improving the performance that is its execution time and also to make the data meaningful is the most important task of this research. The meaningful data is retrieved when we get the best Association rules from both the algorithm. The Execution time for apriori algorithm is better than the FP-Growth algorithm. Therefore for improving the Execution time proposed model is used for both algorithm. It improves the result of FP-Growth and also reduces the execution time for both the algorithm and makes it efficient.

REFERENCES

[1] J. C. S. Anjos, M. D. Assunção, J. Bez, "An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments", *IEEE International Conference on Computer and Information Technology*, pp.199-206, 2015.

[2] X. Wu, X. Zhu, G. Q. Wu, W. Ding, "Data Mining with Big Data", *IEEE transactions on Knowledge and Data Engineering,* Vol.26, pp. 246-250, 2014.

[3] W. Fan, A. Bifet," Mining Big Data: Current Status, and Forecast to the Future", *ACM SIGKDD Explorations*, Vol.14, No. 2, pp.1-5,2013.

[4] C. W. Tsai, C. F. Lai, M. C. Chiang, and L. T. Yang Tsai, ",Data Mining for Internet of Things: A Survey", *IEEE Communications Surveys & Tutorials,* Vol.16, No. 1, pp.77–97.

[5] K. R. Suneetha, R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", *International Journal of Computer Science and Network Security(UCSNS),* Vol. 9, Issue 4, 2009.

[6] F. Z. Benjelloun, A. A. Belfkih, "An Overview of Big Data Opportunities, Applications and Tools", *IEEE Conference on Intelligent Systems and Computer Vision*, 2015.

[7] R. Agrawal , R. Srikant, "Fast algorithms for mining association rules", *In Proc.of the 20th international conference on very large data bases(VLDB),* pp 487–499, 1994.

[8] J. M. Zaki, "Scalable algorithms for association mining", *IEEE Transactions on Knowledge and Data Engineering,* Vol.12, No. 3, pp. 372–390, 2000.

[9] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", *In Proc. of the 20th ACM SIGMOD Record*, Vol.29, No. 2, pp.1-12, 2000.

[10] K. Dharmarajan, M. A. Dorairangaswamy" Analysis of FP-Growth and Apriori Algorithms on Pattern Discovery from Weblog Data", *IEEE International Conference on Advance in ComputerApplication*, 2016.

[11] R. Agrawal and S. Ramakrishnan, "Fast algorithms for mining association rules", *International VLDB Conference*, Vol. 1215, 1994.

[12] R. Agrawal and R. Srikant,"Mining Sequential Patterns", *11th International Conference on Data Engineering,* 1995.

[13] M. Sustersic, D. Mramor, J. Zupan, "Consumer Credit Scoring Models with Limited Data", *Expert Systems with Applications*, Vol. 36, No. 3, pp. 4736-4744, 2009.

[14] K. Rajeswari, "Feature Selection by Mining Optimized Association Rules based on Apriori Algorithm" *International Journal of Computer Applications*, Vol. 119, 2015

[15] K. Chavan, P.Kulkarni, P. Ghodekar," Frequent Itemset Mining for Big Data", *International Conference on Green Computing and Internet of Things,* 2015.

[16] S. V. Nandury, B. A. Begum, "Strategies to Handle Big Data for Traffic Management in Smart Cities", in *International Conference on Advances in Computing, Communications and Informatics*, pp. 21-24, 2016.

[17] R. Sowmya, K. R. Suneetha, "Data Mining with Big Data", *11th International Conference on Intelligent Systems and Control*, pp. 246-250, 2017.

[18] L. Wang, "Heterogeneous Data and Big Data Analytics", in *Automatic Control and Information Sciences*, Vol. 3, No. 1, pp. 8-15, 2017.

[19] M. Nagalakshmi, I.S. Prabha, K. Anil, "Big data implementation of apriori algorithm for handling voluminous data-sets", *International Journal of Engineering & Technology*, pp. 217-220, 2018.

[20] S. E. Bibri, J. Krogstie, "The Big Data Deluge for Transforming the Knowledge of Smart Sustainable Cities: A Data Mining Framework for Urban Analytics", in *SCAM*, 2018.

[21] M. T. Baldassarre, I. Caballero, D. Caivano, B. Rivas, M. Piattini, "From Big Data to Smart Data: A Data Quality Persepective" *International Workshop on Ensemble International Workshop on Ensemble-Based Software Engineering*, 2018.

[22] K. Poonsirivong and C. Jittawiriaynukoon," Big Data Analytics Using Association Rules in eLearning", in *IEEE 3rd International Conference on Big Data Analysis*, 2018