https://dx.doi.org/10.4314/jcsia.v27i1.8

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 27, No. 1 June 2020

COMPARATIVE ANALYSIS OF ASSOCIATION RULE MINING ALGORITHMS IN MARKET BASKET ANALYSIS USING TRANSACTIONAL DATA

A.A. Izang,[1] S.O. Kuyoro,[2] O.D. Alao,[3] R.U. Okoro[4] and [5]O. A. Adesegun[5]
[1,2,3,4,5] Babcock University. Ilishan-Remo, Ogun State, Nigeria.
[1] aaronizang89@gmail.com, [2]afolashadeng@gmail.com, [3]alaool@babcock.edu.ng,
[4]okoror@babcock.edu,ng, [5]oricgoldenfinger@gmail.com

ABSTRACT

Association rule mining (ARM) is an aspect of data mining that has revolutionized the area of predictive modelling paving way for data mining technique to become the recommended method for business owners to evaluate organizational performance. Market basket analysis (MBA), a useful modeling technique in data mining, is often used to analyze customer buying pattern. Choosing the right ARM algorithm to use in MBA is somewhat difficult, as most algorithms performance is determined by characteristics such as amount of data used, application domain, time variation, and customer's preferences. Hence this study examines four ARM algorithm used in MBA systems for improved business Decisions. One million, one hundered and twele thousand (1,112,000) transactional data were extracted from Babcock University Superstore. The dataset was induced with Frequent Pattern Growth, Apiori, Association Outliers and Supervised Association Rule ARM algorithms. The outputs were compared using minimum support threshold, confidence level and execution time as metrics. The result showed that The FP Growth has minimum support threshold of 0.011 and confidence level of 0.013, Apriori 0.019 and 0.022, Association outliers 0.026 and 0.294 while Supervised Association Rule has 0.032 and 0.212 respectively. The FP Growth and Apirori ARM algorithms performed better than Association Outliers and Supervised Association Rule when the minimum support and confidence threshold were both set to 0.1. The study concluded by recommending a hybrid ARM algorithm to be used for building MBA Applications. The outcome of this study when adopted by business ventures will lead to improved business decisions thereby helping to achieve customer retention.

Keywords: Association rule mining, Business ventures, Data mining, Market basket analysis, Transactional data.

1.0 INTRODUCTION

As more and more businesses go electronic, such business organizations would continue to amass large amounts of data from their customers over time [1]. Transactional data streams (TDS) deals with the transactions of the organization and includes data that is captured, when a product is sold or purchased. TDS is therefore usually more volatile than master data (data about business entities that provides context for business transactions) as it is created and changes more frequently as new data streams occur [2]. An enormous amount of transactional

data coming from marketing, fraud detection in surveillance, geological department, issues related to human factor, scientific discovery, data in the medical field, geographical information system, ecosystem amongst others requires in-depth analysis and proper decision making which can be achieved through data mining [3].

Data mining is a method where data analysis is used to discover hidden relationship and patterns, which helps business establishments focus on essential information in data warehouses. Anomaly detection, association rule mining, sequence and path analysis, clustering, classification and prediction, neural networking, forecasting, genetic algorithm, amongst others are the techniques used in data mining. Various fields such as retail marketers, medical field, and banking sector amongst others, need to maintain a large quantity of data in multiple data repository. However, it is not the entire information that will be useful to the user. Hence, it is essential to mine only the valuable information from the massive data available [4]. Knowledge Discovery in a Databases (KDD) is another name for data mining, which determines the correlations, patterns, trends, and anomalies, in these databases that help in making a business decision for the future [5].

Market basket analysis (MBA) is a data mining technique for analyzing the association of items, that frequently occurred together from everyday buying and selling. In order to promote specific items together by controlling the stocks concurrently, and increasing the probability of purchasing, managers in a superstore can place item pair that are associated to the shelf beside them. Thus, giving business organizations a better chance of profit-making through marketing control and controlling the order of goods [6]. The rudimental conception in MBA helps to discover the item pairs associated in the store. The study makes use of purchases from a superstore as the transactional data to solve the problem of choosing the right ARM algorithm for MBA. The aim of studying this is to allow the business organizations to prepare for such occurrences thereby satisfying the customers and also help the business organization improve their revenue.
Association rule mining, is one of the most consequential technique of data mining. The process of knowledge discovery in databases (KDD), includes data selection, pre-processing, transformation, data mining and interpretation. The extraction of hidden patterns of predictive information, and transforming it into an understandable structure, is the major goal of data mining algorithms [7].

Association rules, describe how often items are purchased together when a shopper passes through a point of sale, the contents of his market basket are registered. This results in immensely colossal accumulations of market basket data, that provide information about items that were sold and, in particular, the combinations of items sold together. Thus this is a way of finding unsuspected relationship, for decision making process through the creation of frequent item sets for prediction. Frequent item sets, are those items that occur at least a given number of times for shoppers (referred as minimum support threshold), in the database. Such type of rules, can be helpful for paramount business decisions like promotions, store layout, product pricing, cross-selling amongst others [8].

Despite the numerous advantage offered by using ARM algorithms for MBA, through the help of transactional data, performing market basket analysis still faces big challenges such as time variation during which the transaction occurred, application domain, and customer's preferences. Studies on how to handle transactional data for MBA makes a stationarity assumption by training and testing association rule mining algorithms on transactional dataset gotten from the same population. However, this

deprives proposed association rule mining algorithms of the adaptiveness required to handle the challenges associated with MBA [6]. Hence there is a need to compare and determine the optimal algorithm to be used in MBA systems. Therefore, the aim of this work is to carry out a comparative analysis of Association Rule Mining (ARM) algorithms for market basket analysis of transactional data.

2.0 RELATED WORKS

Author in [9] examined a comparative study of Association Rules Mining Algorithms. The study compared ARM algorithms that uses candidate set generation with the ARM algorithms that do not use candidate set generation during the process of scanning the dataset. A sample test data set was used for the comparison. Apriori, FP-growth and DynFP-growth algorithms were chosen for the comparison. The three algorithm were tested using a test data to consider which one performs better than the other and based on the outcome of the test the algorithms without candidate generation (FP-growth and DynFP-growth) behave much better in terms of performance as compared with the candidate generating algorithms (Derived from Apriori), this is due to the fact that Apriori algorithm can only handle small amount of data (maximum of 50,000 transactions) as discovered from the study.

Author in [10] Surveyed association rule mining as a technique used to extract important patterns from existing information which enables better decision making in an establishment. This paper presented a critical review of various ARM algorithms, comparing each of the algorithms, and considering the merit and demerit of each and applying the outcome in developing applications that conducts MBA. The study shows that choosing an ARM algorithm for MBA depends on the data set size and the application area in MBA that the algorithm will be used. The no free lunch theorem state that no algorithm is guaranteed to outperform others in all domains hence the need for study the algorithms in order to determine their performance. The use of a hybrid ARM algorithm in developing application for MBA was recommended in this study.

Author in [11] carried out a study on MBA using Association rule mining a potent tool for basket analysis, which helps to identify the correlation that exists between items stored in a large database. The researchers studied the buying pattern of customers (Which is the main aim of MBA) through the use of existing data mining algorithms. Furthermore, the study implemented the Apriori algorithm for MBA been one out of many ARM algorithms used to find the frequent itemsets from a given data repository.

Author [12] examined critically Market basket analysis through the use of association rule mining. The study aim was to broaden the concept of ARM for MBA of different items. Researchers collected primary data from retailers and wholesalers, and the data were analyzed using the FP-Growth algorithm with a data mining tool. The FP-Growth ARM algorithm was able to trace the various association rules from the data collected.

Author in [13] surveyed association rule mining in market basket analysis. the advantages of MBA was discovered in this work, some of which focuses on the detection of interesting association relationships between large quantities of business transaction data which helps busineses in catalog design, cross-marketing and various business decision making processes. Also examines customer buying patterns by identifying associations among various items that customers place in their shopping baskets. The identification of such associations can assist retailers expand marketing strategies by gaining insight into which items are frequently purchased by customers. This study acts as a broad area for

the researchers to develop a better data mining algorithm. The outcome of this survey is to discover the existing data mining algorithm for market basket analysis.

Author in [14] examined Market basket analysis in a multiple store environment. The emphasized the importance of MBA over the years has drawn increased research interest because the information obtained from the analysis can be used in forming marketing, sales, service, and operation strategies. The existing methods, however, may fail to discover important purchasing patterns in a multi-store environment, because of an implicit assumption that products under consideration are on shelf all the time across all stores. Therefore in this work, a new method to overcome this weakness was proposed. The empirical evaluation shows that the proposed method is computationally efficient, and that it has advantage over the traditional method when stores are diverse in size, product mix changes rapidly over time, and larger numbers of stores and periods are considered.

3.0 METHODOLOGY

The University superstore transactional datasets which contain 1,112,000 instances, was imported into the Rapid Miner 8.1 and the Tanagra 7.2 data mining tool for the sake of comparing the ARM algorithms to determine the optimal algorithm that will generate more frequent item sets. The reason for using two data mining tools is because most of the tools just have one or two ARM algorithm in them, thereby giving rise to the use of two data mining tool in this work. Each algorithm was trained and tested on different streams of the transactional data, and it was evaluated using the minimum confidence level, minimum support level and execution time as the metrics for evaluating ARM algorithms. The comparison of the output from both data mining tools shows the graphical representation of the performance of the algorithms based on the metrics.

A. Datasets and Arm Algorithms

To carry out the comparative analysis of ARM algorithms for market basket analysis of transactional data, these procedures were followed.

a. Existing ARM algorithms were studied and reviewed in the literature, but for the purpose of this work four algorithms (FP Growth, Apriori, Association Outlier, and Supervised Association rules) were chosen based on its suitability with the data mining tools used.

b. The data used for this work was collected from a University Superstore which comprises of One million, one hundered and twele thousand (1,112,000) transactions, which was divided into the traning and testing data.

c. The data collection process for this study was done using the ETL approach (Data extraction, Data Transformation, and Data loading). At the extraction phase, only items like the snacks (bread and chips), drinks and bread spread item sets were extracted from the superstore transactional data. The data transformation was done using two methods; data cleaning which involves removing unwanted item pairs from the data and data blending which involves putting the data in a way it will be suitable to work with on the data mining tool used in this work.

d. The training dataset after undergoing the ETL approach was then induced with the FP Growth, Apriori, Association Outliers, and Supervised Association rule ARM algorithms using Rapid Miner version 8.1 and the Tanagra version 7.2 data mining tools respectively, to generate the frequent itemsets of previous transactions.

e. The performance of the four ARM algorithms was then tested and compared

using minimum support level, minimum confidence threshold, and execution time as the metrics for evaluating ARM algorithms for market basket analysis of transactional data.

4.0 RESULTS SHOWING THE PERFORMANCE OF THE ASSOCIATION RULE MINING ALGORITHMS

Four association rule mining algorithm (Apriori, Frequent Pattern growth, Association outliers and Supervised Association Rule) algorithm were analyzed by inducing the algorithm on the transactional data set using RapidMiner version 8.1 for the Apriori and FP growth algorithm while the Tanagra 7.2 data mining tool was used for Association Outliers and Supervised Association Rules algorithms. The result of the performance was compared by setting the minimum support threshold and minimum confidence threshold level to 0.1 against the execution time produced by each of the algorithm used.

B. Performance of FP Growth Association Rule Mining Algorithm

The performance of FP growth ARM algorithm using minimum support of 0.1 and confidence level 0.1, shows that most customer's preferred buying bread and drink of the entire transactions analyzed as shown in the result summary. And the time it took for the algorithm to compile was within 0.20 seconds. Appendix A shows a more generalized form of the association rules created. Table 1 shows a summary of the performance of the FP growth algorithm.

Table 1: Summary Result of FP Growth Algorithm

Association rules	Support level	Confidence level	Time taken to execute (s)
BU ENRICHED LARGE BREAD, BU WHOLE WHEAT BREAD --> 7UP PET DRINK 50CL	0.026	0.194	0.20 secs
BU FRUIT MALT BREAD, BU ENRICHED CAKED BREAD LARGE --> C-WAY PEACH 500ML	0.006	0.141	
BU ENRICHED LARGE BREAD; BU ENRICHED CAKED BREAD LARGE --> NUTELLA FERRERO HAZELNUT SPREAD 200G	0.006	0.114	
BU ENRICHED CAKED BREAD MEDIUM, BU ENRICHED LARGE BREAD --> FANTA DRINK 50CL	0.011	0.013	

C. Performance of Apriori Association Rule Mining Algorithm

The performance of the Apriori ARM algorithm using minimum support threshold of 0.1 and confidence level 0.1, shows that most of the customer's preferred buying bread and drink of the entire transactions analyzed. And the time it took for the algorithm to compile was within 0.26 seconds. The summary of Apriori algorithm performance is shown in Table 2.

Table 2: Summary Result of Apriori Algorithm

Association rules	Support level	Confidence level	Time taken to execute (s)
BU ENRICHED LARGE BREAD, BU WHOLE WHEAT BREAD --> 7UP PET DRINK 50CL	0.062	0.014	0.26 secs
BU FRUIT MALT BREAD, BU ENRICHED CAKED BREAD LARGE --> C-WAY PEACH 500ML	0.019	0.022	
BU ENRICHED LARGE BREAD, BU ENRICHED CAKED BREAD LARGE --> NUTELLA FERRERO HAZELNUT SPREAD 200G	0.009	0.021	
BU ENRICHED CAKED BREAD MEDIUM, BU ENRICHED LARGE BREAD --> FANTA DRINK 50CL	0.022	0.031	

D. Performance of Association Outliers Algorithm

The performance of the Association Outliers algorithm using minimum support threshold of 0.1 and confidence level 0.1, shows that most customer's preferred buying bread and drink of the entire transactions analyzed as shown in the result summary. And the time it took for the algorithm to compile was within 0.30 seconds. The summary of Association Outliers algorithm performance is shown in Table 3.

Table 3: Summary Result of Association Outliers Algorithm

Association rules	Support level	Confidence level	Time taken to execute (s)
BU ENRICHED LARGE BREAD, BU WHOLE WHEAT BREAD --> 7UP PET DRINK 50CL	0.054	0.294	0.30 secs
BU FRUIT MALT BREAD, BU ENRICHED CAKED BREAD LARGE --> C-WAY PEACH 500ML	0.026	0.412	
BU ENRICHED LARGE BREAD, BU ENRICHED CAKED BREAD LARGE --> NUTELLA FERRERO HAZELNUT SPREAD 200G	0.034	0.421	
BU ENRICHED CAKED BREAD MEDIUM, BU ENRICHED LARGE BREAD --> FANTA DRINK 50CL	0.032	0.342	

E. Performance of Supervised Association Rule Algorithm

The performance of the supervised association rule algorithm using minimum support threshold of 0.1 and confidence level 0.1, shows most customer's preferred buying bread and drink of the entire transactions analyzed as shown in the result summary. And the time it took for the algorithm to compile was within 0.35 seconds. The summary of supervised association rule algorithm performance is shown in Table 4.

Table 4: Summary Result of Supervised Association Rules Algorithm

Association rules	Support level	Confidence level	Time taken to execute (s)
BU ENRICHED LARGE BREAD, BU WHOLE WHEAT BREAD --> 7UP PET DRINK 50CL	0.064	0.221	0.35 secs
BU FRUIT MALT BREAD, BU ENRICHED CAKED BREAD LARGE --> C-WAY PEACH 500ML	0.066	0.212	
BU ENRICHED LARGE BREAD, BU ENRICHED CAKED BREAD LARGE --> NUTELLA FERRERO HAZELNUT SPREAD 200G	0.043	0.231	
BU ENRICHED CAKED BREAD MEDIUM, BU ENRICHED LARGE BREAD --> FANTA DRINK 50CL	0.032	0.331	

F. Comparison of the Classification Algorithm Performance

Based on the performance of the four algorithms discussed earlier, there is a need to compare their performances so as to aid the decision of selecting the optimal algorithm based on the following benchmarks (support level, confidence level, and execution time of the algorithms).

G. Comparison Based on Support and Confidence Level

The outcome of both Support and Confidence level shows that the FP growth and Apriori Algorithms performed better because it fall below the minimum support and confidence threshold of 0.1 for the association rule with the lowest threshold for each of the algorithms. Table 5 shows summary of the algorithms with their respective lowest minimum support and confidence threshold level.

Table 5: Comparison based on Minimum Support and Confidence Threshold Level

Algorithm	Minimum Support threshold	Minimum Confidence threshold
FP Growth	0.011	0.013
Apriori	0.019	0.022
Association Outliners	0.026	0.294
Supervised Association Rules	0.032	0.212

Hence, from the comparison of the performance of the four algorithms based on minimum support and confidence threshold level, the result signified that FP Growth and Apriori Algorithm outperformed the other two algorithm generating association rules lower than the minimum support and confidence threshold as shown in Figure 1.

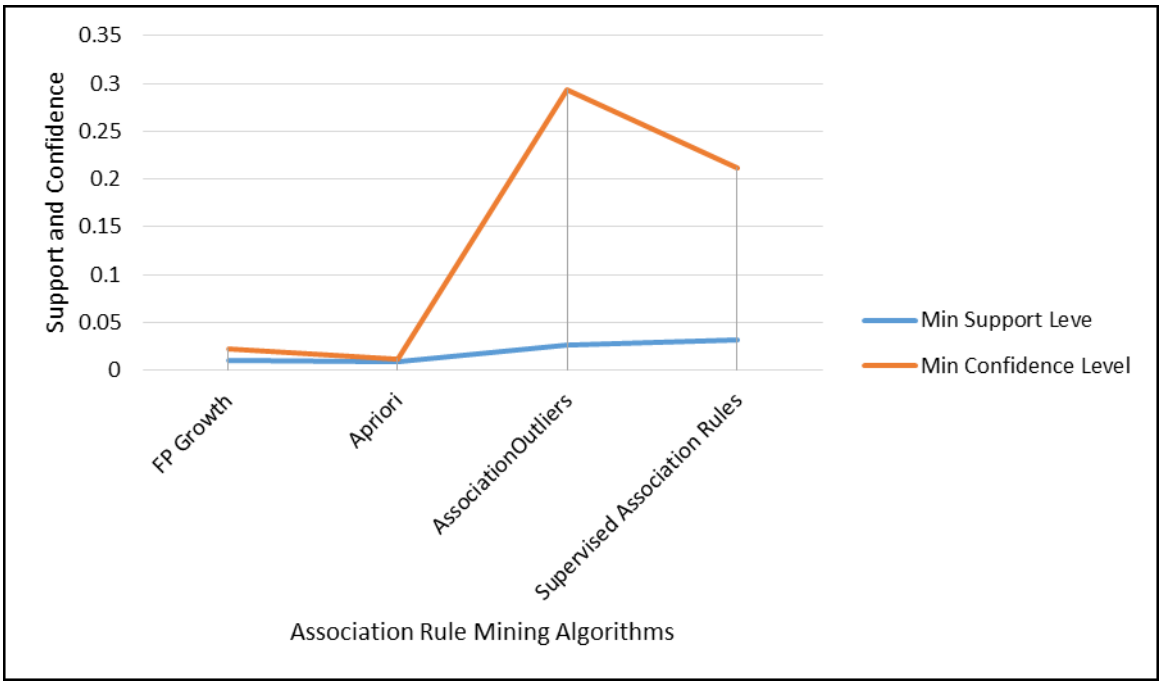

Figure 1: Comparison Based on Support and Confidence Level

H. Comparison Based on Execution Time and Support Level

The outcome of both Execution Time and Support level shows that the FP growth and Apriori Algorithms performed better because it fall below the minimum support threshold of 0.1 for the association rule with the lowest threshold for each of the algorithms. Table 6 shows summary of the algorithms with their respective lowest minimum support and confidence threshold level.

Table 6: Comparison based on Minimum Support and Confidence Threshold Level

Algorithm	Minimum Support threshold	Execution time (Sec)
FP Growth	0.011	0.20
Apriori	0.019	0.26
Association Outliners	0.026	0.30
Supervised Association Rules	0.032	0.35

Hence, from the comparison of the performance of the four algorithms based on minimum support and execution time, the result signified that FP Growth and Apriori Algorithm outperformed the other two algorithm generating association rules lower than the minimum support with lower execution time as shown in Figure 2.

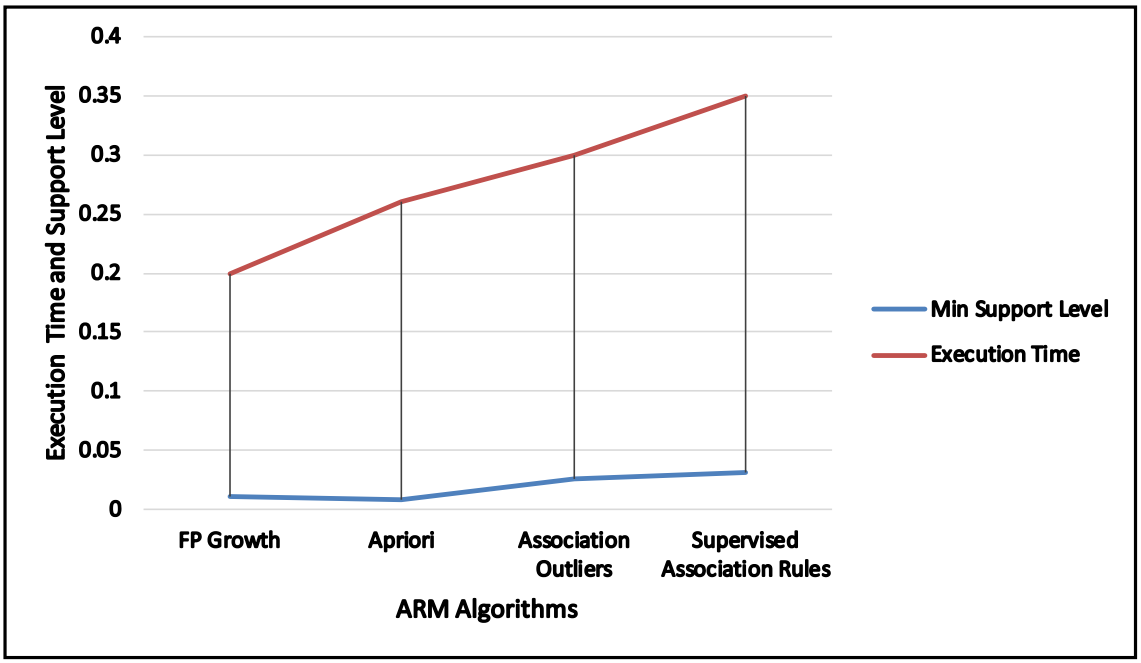

Figure 2: Comparison Based on Support Level and Execution Time

I. Comparison Based on Execution Time and Confidence Level

The outcome of both Execution Time and Confidence level shows that the FP growth and Apriori Algorithms performed better because it fall below the minimum confidence threshold of 0.1 for the association rule with the lowest threshold for each of the algorithms. Table 7 shows summary of the algorithms with their respective lowest minimum confidence threshold level and execution time.

Table 7: Comparison based on Execution Time and Confidence Threshold Level

Algorithm	Minimum Confidence threshold	Execution time (Sec)
FP Growth	0.013	0.20
Apriori	0.022	0.26
Association Outliners	0.294	0.30
Supervised Association Rules	0.212	0.35

Hence, from the comparison of the performance of the four algorithms based on minimum confidence and execution time, the result signified that FP Growth and Apriori Algorithm outperformed the other two algorithm generating association rules lower than the minimum confidence threshold with lower execution time as shown in Figure 3.

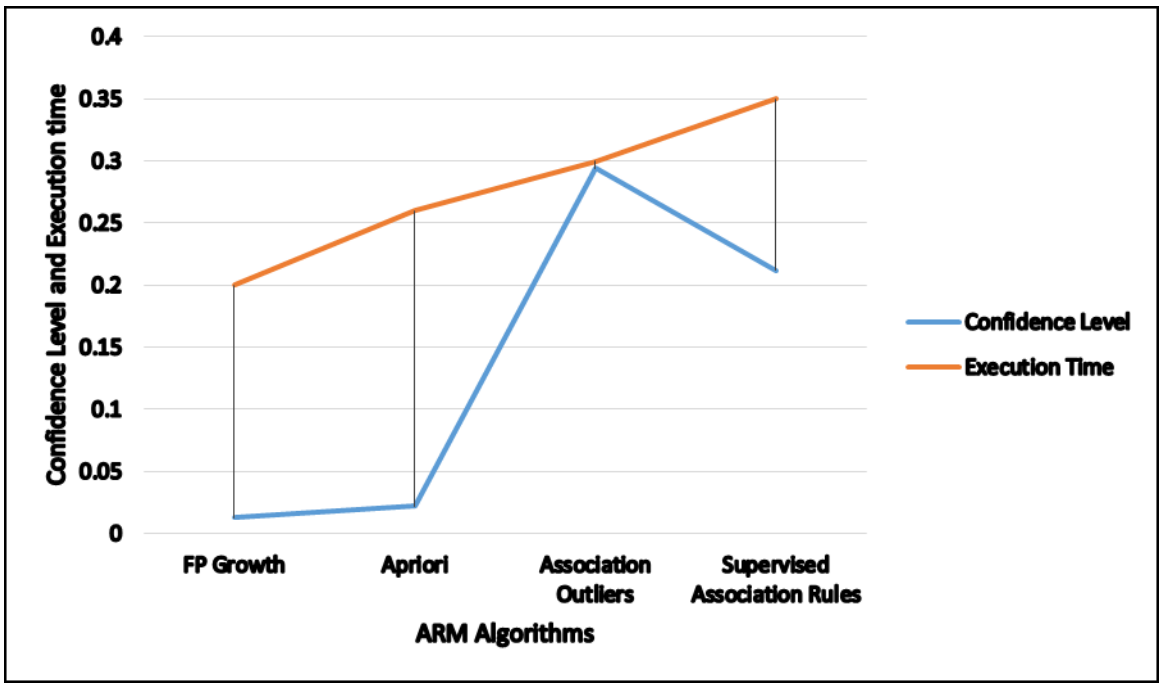

Figure 3: Comparison Based on Confidence Level and Execution Time

4.0 DISCUSSION OF FINDINGS

The results from this study showed that FP growth and Apriori ARM algorithms performed better than Association Outliers and Supervised Association Rule in generating the frequent itemsets which are a representation of the customers buying pattern. Examining the result of the comparison further, it will be discovered that the FP growth and Apriori algorithm has confidence level and support threshold close to the minimum threshold of 0.1. While the Association Outliers and Supervised Association Rule has confidence level and support threshold far apar to the minimum threshold of 0.1. Considering the fact that in evaluating ARM algorithms, algorithms whose confidence level and support threshold are closer to the minimum threshold will perform better in a MBA system. The study was able to prove the idea of using more than one ARM algorithm in MBA as a better solution to undertsanding customer preferences in studying their purcaheses. Therefore the study also discovered that in other to solve the challenges of choosing an appropriate ARM algorithm in MBA, it is imperative to use an enormous amount of transactional data of customers over a very long period. The data will help in predicting the customers buying pattern thereby solving the problem of customer preference, time variation and application domain in MBA. The outcome of this study will help business ventures, retail marketers and business organizations know the preferences of their customers and the items they purchase alongside each other. These will help the business stock their shop with such items in order to enhance customer retention and also improve their revenue.

5.0 CONCLUSION AND RECOMMENDATIONS

The study of Market Basket Analysis (MBA) through the use of Association Rule Mining algorithms helps retail marketer study customers consumption pattern. These is in order to see the relationships that exist

between the items they purchase during a single visit to the superstore.

Most of the past studies have recorded great success in trying to overcome these problems, however, algorithms developed still needs improvement especially in the area of using a large amount of data on the algorithm for analysis, customer preferences and time variation. These work tried to critically study ARM algorithms, and compared some of the algorithm with the help of a large transactional data, over a period of 24 months to determine their performance. Furthermore, this work, in general, provided a more advanced strategy for managing drift in the buying concept of customers, hence help improve the customers shopping experience as well as help the superstore achieve customer retention. This can enhances monitoring the buying pattern of the customers thereby generating more revenue for the superstore in Market basket analysis. However, Based on the comparison of the ARM algorithms, this work came up with a bullet prove solution that using one algorithm is not sufficient enough to improve the performance of a business organization as only one algorithm cannot give the business a required adequate recommendation. This work thereby recommends a hybrid algorithm that will combine the attribute of both algorithm in one be used. Using this hybrid algorithm will provide the required result needed to build a MBA system will be achieved and this in turn will be used to implement an intelligent market basket system that can be used by business organizations.

REFERENCES

[1] Ojugo, A.A, & Eboka, A.O (2018). Modeling the Computational Solution of Market Basket Associative Rule Mining Approaches Using Deep Neural Network. Digital Technologies, 3(1), 1-8.

[2] Parneeta, D., & Bhatia, M.P. (2017). A two ensemble system to handle concept drifting data streams: recurring dynamic weighted majority. *International Journal of Machine Learning and Cybernetics* 10(4), 8 19, DOI: 10.1007/s13042-017-07389

[3] Thirunavukkarasu, K., Digvijay, S., Mohd, A., Ajay, S. S. (2016). Data Mining Techniques in Cloud Computing: A Survey. *International Journal of Recent Trends in Engineering & Research (IJRTER), 2* (3), 288-296.

[4] Manpreet, K & Shivani, K. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science* 85. 78-85. doi:10.1016/j.procs.2016.05.180.

[5] Dhanabhakyam, M., & Punithavalli, M. (2011). A Survey on Data Mining Algorithm for Market Basket Analysis. *Global Journal of Computer Science and Technology*, XI(11), 13 21. Retrieved Feburary 8, 2017.

[6] Izang, A., Goga, N., Kuyoro, S.,Alao, O., Omotunde, A., & Adio, A. (2019). Scalable Data Analytics Market Basket Model for Transactional Data Streams (IJACSA) *International Journal of Advanced Computer Science and Applications*, 10(10), 61-68.

[7] Nidhi, M., Nikhilendra, K. P., & Pankaj, A. (2016). Market Basket Analysis using Association Rule Learning. *International Journal of Computer Applications Recent Trends in Future Prospective in Engineering & Management Technology, 21(9),* 213-220.

[8] Deepa, S. D. (2014). A Novel Approach for Association Rule Mining using Pattern Generation. *International Journal of Information Technology and Computer Science, 11(3),* 59-65. DOI: 10.5815/ijitcs.2014.11.09.

[9] Corneli, G., Robert, G., & Stefan, H. (2003). A Comperative Study of Association Rules Mining Algorithms. IV(2), 40-50.

[10] Izang, A., Goga, N., Kuyoro, S., & Adetunji, O. (2017). An Overview Of Association Rule Mining (ARM) Algorithms for Market Basket Analysis (MBA). *Journal of Research in Engineering and Applied Sciences, II*(4), 132-140. Retrieved from www.jreas.com.

[11] Phani, P., & Murlidher, M. (2013). A Study on Market Basket Analysis Using a Data Mining Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, III(6), 361-363. Retrieved from www.ijetea.com.

[12] Maheshwari, N., Pandey, N. K., & Agarwal, P. (2016). Market Basket Analysis using Association Rule Learning. *International Journal of Computer Applications*, 0975-8887. Retrieved Feburary 12, 2017, from www.ijcaonline.org

[13] Gupta, S., & Mamtora, R. (2014) A Survey on Association Rule Mining in Market Basket Analysis, *International Journal of Information and Computation Technology*. 4(4), 409-414, ISSN 0974-2239.

[14] Chena, Y.L., Tang, K., Shen, R.J., & Hu, Y.H. (2004) Market basket analysis in a multiple store environment, *Decision Support Systems* 40 (2005) 339-354, Available online at www.sciencedirect.com, Published by Elsevier.com