International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Analyzing the Performance of ECLAT Algorithm for Large Datasets by Comparing K-means and Gaussian Mixture Model

Nandar Lin[a]*, Thanda Win[b]

[a,b]Computer Engineering and Information Technology, Yangon Technological University,Yangon, Insein, 11012, Myanmar

[a]Email: nandarlin711l@gmail.com

[b]Email: thanda80@gmail.com

Abstract

Frequent Itemset Mining (FIM) is a technique that transforms historical data into useful information by identifying beneficial patterns. The ECLAT method uses depth-first search to intersect the transaction ID sets with the corresponding k^{th} item sets in order to calculate the support items. While searching for the best-selling products, ECLAT uses a lot of memory and processing time due to the enormous number of transaction ID sets. To overcome these problems, the clustering method combines with the ECLAT algorithm to retrieve the support items. Description elements 100,000 to 400,000 were used to retrieve the support items of the most popular selling goods. For the K-means clustering approach, the optimal value of k is 8 clusters according to the 0.59 silhouette value. For the Gaussian Mixture Model, the ideal value of k is 14 clusters based on a 0.59 silhouette score value between 100,000 and 400,000 data items. After clustering the same product items, the ECLAT algorithm retrieves the support items by applying a minimum support value of 0.00001 in this investigation. According to the experimental results, the Gaussian Mixture Model not only offers more flexibility for clustering the same items but also reduces the memory usage and execution times. The outcomes of this investigation indicate that the Gaussian Mixture Model provides more efficient enhancement of the performance of the ECLAT algorithm than the K-means algorithm.

Keywords: Frequent Itemset Mining; Support Items; ECLAT; K-means; Gaussian Mixture Model.

Received: 2/27/2025
Accepted: 4/25/2025
Published: 5/6/2025

----

* Corresponding author.

1. Introduction

A rule-based machine learning technique called association rule mining [1] is used to find major relationships between items that represent market basket items, user actions in intrusion detection, stock analysis, bioinformatics, medical diagnosis, and business decisions, as well as the consecutive clicks made by the corresponding users in click streams. Frequent Itemset Mining (FIM), initially presented in [2], is an important process in association rule mining. The goal of this unsupervised data-mining method is to identify items that frequently occur together in transactions inside a transaction dataset. Since it searches for each possible combination of frequently occurring elements that are also often occurring in the transaction dataset, it is a very computationally and spatially demanding task.

The most resource-demanding part of the FIM algorithm is the support counting of itemset from the transaction dataset. In a vertical dataset format of ECLAT algorithm, ECLAT algorithm determines that itemset {A, B, C} is also frequent, it needs to cross the TIDs of itemset {A, B} and item C, hence the result of this intersection is also stored. Without storing this intermediate result, ECLAT algorithm would perform two intersections, meaning that the result of A and B intersecting with C. This raises memory usage even as it saves computations. The majority of the time and space needed by a FIM method is accounted for by these transaction IDs and the intersection operations performed on them.

There are two primary causes of the FIM algorithms' performance bottleneck: 1) the computation of itemset support, a regularly performed and extremely processing-intensive function; and 2) the need for large amounts of memory to store transaction data. For large datasets, the amount of time required to support computations may rapidly exceed acceptable bounds. The problem's complexity is increasing due to the rapid growth of transaction data created by decades of information technology advancements, which presents enormous difficulties for accurate FIM algorithms.

To solve the problem of more memory requirement, clustering technique is used in this paper. The technique of clustering involves dividing data into many clusters or groups so that the degree of similarity between data in one cluster is at its highest and that of similarity between clusters is at its lowest. This research not only solves the memory requirement and execution time but also analyzes the ECLAT algorithm's performance by comparing K-means and Gaussian Mixture Model.

2. Related Work

Dr. G. Naga Chandrika, G. Varshith, N. Bhargav Reddy, and G. Gurubrahmaiah proposed [4] that customer segmentation is a marketing strategy to improve customer relationships and loyalty. However, many companies lack a segmentation system to understand customer types and measure their value. This study aimed to identify customer criteria using RFM values, based on clustering methods like K-Means Clustering and Gaussian mixture algorithms. The company's K-Means algorithm and Gaussian mixture models test indicated cluster 3 as the optimal promotional medium for loyal customers, with a significant difference value of Sum Square Error of 2.7630 and the Silhouette index of 0.7210.

According to this research [5], M. Hafidh Raditya, Indwiarti, and A. Atiqi Rohmawati are proposed that these research uses the Gaussian Mixture Model-Based Clustering Method with the Expectation-Maximization algorithm to segment house prices in Jakarta into low-profile, mid-profile, and high-profile houses. The results show that house prices are becoming more varied based on house parameters like area, location, and number of rooms. The study aims to optimize Gaussian Mixture Model (GMM)parameters and analyze house price segmentation results using this method. The dataset used is house prices data from the www.olx.co.id website, obtained in January 2022.

In [6], N. P. Dharshinni, H. Mawengkang, and M. K. M. Nasution proposed that medicine is one of the items needed by sick society; the high influence of medicine on service and the economy in hospitals requires mapping and planning the optimal need for medicines according to the conditions because 50%-60% of hospital income is sourced from medicine sales. The purpose of this study was to find patterns of doctors' prescription medicine association with sales data using an apriori algorithm based on data grouping using a k-means algorithm. The results of the experiments show that medicine prescription data with medicine sales have significant differences so that the data can’t be used as materials for medicine planning. This is due to some indication of one of the unavailability of medicine caused by mapping inaccuracy so that the planning of medicine requirements is not optimal. The results of this analysis can be used as input materials in decision-making, so the planning needs of medicines can be in accordance with the development of patient disease patterns. The study reveals significant differences between sales data and patient medical data, with 35% of doctor research data not involving sales transactions. This information is crucial for decision-making in medicine planning. The k-means clustering method significantly influences association rules and computation time.

In [7], C.P. Ezenkwu, S. Ozuomba, C. Kalu presented that big data and machine learning have led to automated customer segmentation. The paper presents a MATLAB implementation of the k-Means clustering algorithm for customer segmentation, based on dataset of 100 training patterns acquired from a retail business. The features are the average amount of goods purchased by customer per month and the average number of customer visits per month. Four customer clusters were identified with 95% accuracy, labeled as High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers Regular-Visitors (LBRV), and Low-Buyers-Irregular-Visitors (LBIV).

This paper [8] compared the effectiveness of Support Vector Machine (SVM) and Artificial Neural Network (ANN) models for predicting option prices in the financial market. Both models are tested using a publicly available dataset, SPY option price-2015. The study compares the performance of the SVM and ANN models using Principal Component Analysis (PCA) data and partitioning the dataset into training and test sets to avoid overfitting. Results show ANN performs better than SVM and predicts option prices accurately.

In this research, combining the K-means and ECLAT algorithms is used to retrieve the support items of best-selling products. Firstly, the retail dataset for UK consumers is grouped into the same clusters by applying the K-means algorithm. Then grouping data is used by the ECLAT algorithm to produce support for top-selling items. Then the Gaussian Mixture Model clusters the data items into the same clusters, and the ECLAT algorithm is used to produce the support items from top-selling products.

3. Research Method

In this paper, support items are retrieved by combining the ECLAT algorithm with clustering approaches in order to solve the memory need. Firstly, K-means algorithm group the same items and ECLAT algorithm produce the support items from the clustering outcomes. Secondly, Gaussian Mixture Model cluster the top selling items and then ECLAT algorithm retrieve the support items from top selling items.

3.1. K-means Algorithm

One of the most significant algorithms for data grouping is K-means. Since the data points belong to a single group after classification, clustering can be thought of as grouping. Conversely, the data is simultaneously assigned to several groups by the gentle clustering methods. K-means automatically modifies the center of each group based on the distance to the data points until the algorithm converges, taking into account the number of clusters. As an aside, there are various kinds of distances; Euclidean distance, or the geometric separation between two data points, is used by K-means [8]. The K-means algorithm is used to cluster the same items into different groups according to the Euclidean distance. Hugo Steinhaus and James MacQueen raised the K-means algorithm for the first time. It can generally be divided into three primary sections [9].

- Centroids Initialization: To choose k observations at random from the dataset in order to initialize the centroids.
- Assigning Step: Assign each observation to the closest cluster (the one where the mean distance between the observation and the cluster is the least when compared to other   clusters).
- Update Step: For every new cluster, recalculate the mean to be its centroid [9].

The K-means algorithm is described as follows:

1. Initialize the k centroids
2. Calculate the distances between each observation and each centroid.
3. Allocate each observation to the nearest centroid.
4. Recalculate the mean to be the centroid of each new cluster.
5. Repeat 2 to 4 until convergency happens
6. Repeat step 1 to step 5.

$$\text{Euclidean Distance, d} = \sqrt{(x2 - x1)^2 + (y2-y1)^2} \qquad (1)$$

3.2. Silhouette Score

Silhouette Score is a tool for evaluating the appropriateness of clustering results. The Silhouette Score measures a data point's uniqueness from other clusters and how well it fits into its designated cluster. It assists in determining if the clusters are well-separated and internally homogeneous by measuring the cohesion and separation of data points inside clusters. The Silhouette coefficient is a value between -1 and 1, where higher values indicate a better clustering [10].

The Silhouette score for each data point i is calculated as follows

$$\text{Silhouette Score (i)}=\max(a_i,b_i)/(b_i-a_i) \qquad (2)$$

Where, a_i= The average distance of i to all other data points in the same cluster (intra-cluster distance)

$b_{i=}$ The average distance of i to all data points in the nearest cluster (inter-cluster distance)

3.3. Sum Square Error

Error Sum of Squares (SSE) is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. The within-cluster sum of squares is a measure of the variability of the observations within each cluster. In general, a cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares.

3.4. Elbow Method

When using K-means clustering, the Elbow Method is a visual method for figuring out the optimal "K" (number of clusters). The method involves computing the Within-Cluster Sum of Squares (WCSS), which is the sum of the squared distances between each cluster center and each data point. But eventually, the rate of decrease slows down and increasing K no longer results in a statistically significant reduction in WCSS. The elbow is the term used to describe this place. The elbow graph shows the within-cluster-sum-of-square (WCSS) values on the y-axis corresponding to the different values of K (on the x-axis). The optimal K value is the point at which the graph forms an elbow.

3.5. Gaussian Mixture Model

In machine learning, the Gaussian mixture model (GMM) is a probabilistic model that groups data by finding similarities and differentiating between them. It describes each Gaussian distribution using a mean and covariance matrix. This is how the GMM algorithm works [11]:

- Initialize phase: The parameters of Gaussian distributions should be initialized. (means, covariances, and mixing coefficients).
- Expectation phase: Calculate the likelihood that each of the Gaussian distributions was used to create each data point.
- Maximization phase: To re-estimate the Gaussian distribution parameters, use the probabilities discovered in the expectation phase.
- Final phase: Repeat the second and third stages, to get the parameters to converge.

3.6. Likelihood

The observed data under a hypothetical situation is represented by a likelihood. The Gaussian log-likelihood function calculates the likelihood of all measurements following a Gaussian distribution for indicating the goodness of fit for a model. The model is better if the value is small negative value. It is possible for the log probability to fall between negative and positive infinity.

3.7. ECLAT Algorithm

The "Equivalence Class Clustering and Bottom-Up Lattice Traversal" (ECLAT) algorithm is a depth-first search strategy for locating frequently occurring item sets. Step by step explanation are as follows

- Transaction Database: ECLAT begins with a transaction database, in which an item is represented by each column and a transaction by each row. Either a 1 (signaling the presence of an item in a transaction) or a 0 (signaling its absence) is present in every cell.
- Itemset Generation: Initially, ECLAT generates 1 item sets, which are lists of single things. Each item in the database is scanned once to determine its support (frequency).
- Building Equivalence Classes: ECLAT groups transactions with similar items in their 1-itemsets to create equivalence classes. Equivalence classes reduce the number of potential itemset combinations to consider.
- Recursive Search: ECLAT combines smaller item sets to search bigger ones recursively. It does this by taking the intersection of equivalence classes of items. This step is similar to the join operation in the Apriori algorithm.
- Pruning: Similar to Apriori, ECLAT reduces the search space by pruning infrequent itemset at each stage. If an itemset support falls below a predefined minimum support threshold, is eliminated.
- Repeat: Iterative recursive search and trimming procedures are used to find every frequently occurring itemset in the dataset [12].

4. Proposed Method

Figure 1 displays the proposed method's flow diagram. First, the retail dataset for the UK was obtained by downloading the online. The retail dataset contains negative values for the quantity of goods and additional space for the description. The first step of data cleaning involves removing extraneous spaces from the description. The next step is to preprocess the data so that only item descriptions are obtained and any negative quantity values are removed. Following preprocessing, the clustering algorithm applies cleaning data to group the same elements. Based on 0.59 silhouette score value, the K-means algorithm produces 8 clusters and Gaussian Mixture Model creates 14 clusters for the best-selling products, according to the items ranging from 100,000 to 400,000. The ECLAT algorithm then uses the output cluster items to create support items. The ECLAT algorithm sets a minimum support of 0.00001 in order to extract support items. The lower the minimal support can search the more support items for large dataset.

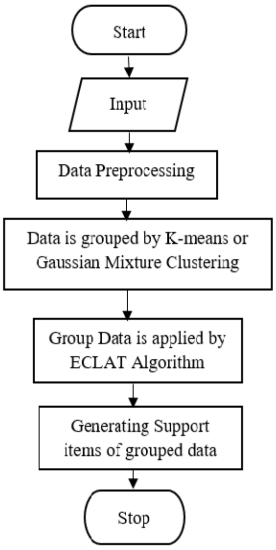
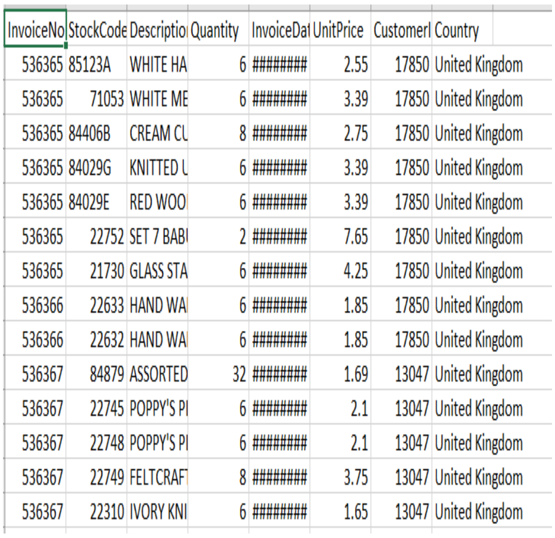

Start
Input
Data Preprocessing
Data is grouped by K-means or Gaussian Mixture Clustering
Group Data is applied by ECLAT Algorithm
Generating Support items of grouped data
Stop

Figure 1: flow diagram of proposed method

5. Implementation

The downloadable Online Retail Dataset was used in the current investigation. A large number of the company's customers are retailers. A variety of handmade gifts suitable for every celebration make up the company's primary product line. The dataset consists of eight features: UnitPrice, Customer, Country, Invoice number, StockCode, Description, Quantity, and Invoice date. There are 541910 records in the collection primarily. There are 38 distinct countries represented in this dataset. The dataset is sampled in Figure2.



InvoiceNo	StockCode	Descriptio	Quantity	InvoiceDat	UnitPrice	CustomerI	Country
536365	85123A	WHITE HA	6	########	2.55	17850	United Kingdom
536365	71053	WHITE ME	6	########	3.39	17850	United Kingdom
536365	84406B	CREAM CU	8	########	2.75	17850	United Kingdom
536365	84029G	KNITTED U	6	########	3.39	17850	United Kingdom
536365	84029E	RED WOO	6	########	3.39	17850	United Kingdom
536365	22752	SET 7 BAB	2	########	7.65	17850	United Kingdom
536365	21730	GLASS STA	6	########	4.25	17850	United Kingdom
536366	22633	HAND WA	6	########	1.85	17850	United Kingdom
536366	22632	HAND WA	6	########	1.85	17850	United Kingdom
536367	84879	ASSORTED	32	########	1.69	13047	United Kingdom
536367	22745	POPPY'S PI	6	########	2.1	13047	United Kingdom
536367	22748	POPPY'S PI	6	########	2.1	13047	United Kingdom
536367	22749	FELTCRAF	8	########	3.75	13047	United Kingdom
536367	22310	IVORY KNI	6	########	1.65	13047	United Kingdom

Figure 2: sample dataset [13]

6. Results and Discussions

The research performs the Intel® CoreTM i7-9700 CPU running at 3.00GHz, 32GB of RAM, a 3TB hard drive, and an Aorus XTREME Geforce RTX 2080Ti GPU. This study tests between 100,000 and 400,000 description elements in the suggested system for UK customers. Initially, it uses the K-means methods to cluster the data items between 100,000 and 400,000. After clustering the same data items, the ECLAT method extract the support items by setting the minimum support 0.00001. As per the experimental findings, the best clustering result is eight clusters, with a sum square error below 0. 7 and a silhouette score of 0. 59, within the range of 100,000 to 400,000. Gaussian Mixture Model groups the description elements into same groups and then ECLAT algorithm retrieves the support items by adjusting 0.00001 minimum support. Based on the analysis, the log probability value of -2.42 and the silhouette score value of 0.59 indicate that the optimal number of k is 14 clusters. This research compares the two clustering methods of unsupervised classification K-means and GMM (Gaussian Mixture Model). According to the experimental results, GMM can more cluster the description elements than K-means on the same silhouette value.

6.1. Optimal Cluster Determination for K-means Algorithm

Table 1 shows the number of clusters and silhouette score value between 100,000 and 400,000 data items. The K-means algorithm tests k values ranging from 3 to 15 to cluster the same products. The optimal cluster is 8 clusters according to the silhouette score value of 0.59, ranging from 100000 to 400000.

Table 1: Silhouette Score Values and Number of Clusters

Data Amount	No: of clusters	Silhouette Score
100,000	8	0.586
200,000	8	0.591
300,000	8	0.588
400,000	8	0.594

6.2. Optimal Cluster Determination for Gaussian Mixture Model

Table 2 indicates the number of clusters and silhouette score value between 100,000 and 400,000 data items. Gaussian Mixture Model takes k values between 3 and 15. The ideal value of k is 14 clusters based on the silhouette score value of 0.59.

Table 2: Silhouette Score Values and Number of Clusters

Data Amount	No: of clusters	Silhouette Score
100,000	14	0.588
200,000	14	0.589
300,000	14	0.591
400,000	14	0.587

The outcomes from the above two tables describe the different optimal cluster values based on the same value of 0.59 silhouette score value. For the K-means algorithm, the ideal value of k is 8 clusters. For Gaussian Mixture Model, the best cluster value of k is 14. So, Gaussian Mixture Model provides more clustering data items than the K-means algorithm. Because clusters are distributed according to a Gaussian (normal) distribution. Because of this supposition, Gaussian Mixture Model may simulate clusters of various sizes and shapes, including elliptical clusters. However, K-means assumes that clusters are spherical and of equal size. It uses the Euclidean distance to assign points to the nearest cluster.

6.3. Producing Support Items by Combining K-means and ECLAT Algorithm

After clustering the product items by using the K-means algorithm, the ECLAT algorithm seeks support items from 100,000 product items to 400,000 by setting the minimum support value to 0.00001, as shown in Table 3. It can produce more support items for a larger amount of data items because a smaller minimum support value can search for more support items.

Table 3: Support Items between 100,000 and 400,000 Data Items

Data Amount	100,000	200,000	300,000	400,00
OVAL WALL MIRROR DIAMANTE	0.10713	0.116475	0.110810	0.190075
50’s CHRISTMAS GIFT BAG LARGE	0.12233	0.132820	0.190363	0.108060
4 PURPLE FLOCK DINNER CANDLE LOVE LONDON	0.1053	0.096065	0.114013	0.092167
MINI BACKPACKSET	0.11519	0.113185	0.129763	0.119017
SET 2 TEA TOWELS I LOVE LONDON	0.13795	0.116575	0.121047	0.128085
RED SPOT GIFT BAG LARGE	0.17309	0.121490	0.091647	0.113837
NINE DRAWER OFFICE TIDY	0.12044	0.182045	0.128870	0.117693
DOLLY GIRL BEAKER	0.11854	0.121345	0.113487	0.131065

6.4. Producing Support Items by Combining Gaussian Mixture Model and ECLAT Algorithm

After clustering the product items by using the Gaussian Mixture Model, the ECLAT algorithm produces support items from 100,000 product items to 400,000 by setting the minimum support value to 0.00001, as shown in Table 4. It can provide more support items for a larger amount of data items because a smaller minimum support value can search for more support items.

Table 4: Support Items between 100,000 and 400,000 Data Items

Data Amount	100,000	200,000	300,000	400,00
OVAL WALL MIRROR DIAMANTE	0.07427	0.123065	0.054237	0.080162
50's CHRISTMAS GIFT BAG LARGE	0.5058	0.064200	0.039307	0.097545
4 PURPLE FLOCK DINNER CANDLE	0.06401	0.070680	0.068667	0.114123
LOVE LONDON MINI BACKPACKSET	0.06121	0.049120	0.061863	0.071015
SET 2 TEA TOWELS I LOVE LONDON	0.10089	0.087455	0.047493	0.086750
RED SPOT GIFT BAG LARGE	0.06401	0.070680	0.068667	0.114123
NINE DRAWER OFFICE TIDY	0.06269	0.104885	0.050507	0.046657
DOLLY GIRL BEAKER	0.07427	0.123065	0.054237	0.080162
SPACE BOY BABY GIFT SET	0.05206	0.056400	0.108873	0.099428
TOADSTOOL BEDSIDE LIGHT	0.06120	0.066850	0.076407	0.041245
Boombox iPod Classic	0.05350	0.53305	0.53311	0.55605
*USB Office Mirror Ball	0.04874	0.037010	0.072053	0.057585
TRELLIS COAT RACK	0.07000	0.055200	0.062273	0.067748
12 COLOUR PARTY BALLONS	0.05350	0.53305	0.53311	0.55605

6.5 Memory Usage and Execution Times for Combining K-means and ECLAT Algorithm

The performance of the proposed method was evaluated variety of data amount. For 100,000 items, the memory usage was 217.672 MB, and the execution time was 1.17 seconds. Memory consumption went up a little to 226.953 MB when the data was doubled to 200,000 items, while the execution time ascended to 1.41 seconds. Memory usage increased to 236.430 MB when scaling up to 300,000 items, and execution time decreased to 1.33 seconds. Memory usage peaked at 249.105 MB at 400,000 items, which corresponds to an execution time of 1.69 seconds as shown in Table 5.

Table 5: Comparison of Memory Usage and Execution Time between 100,000 and 400,000 Data Items

Data Items	Memory Usage (MB)	Execution Time(s)
100000	217.672	1.17
200000	226.953	1.41
300000	236.430	1.33
400000	249.105	1.69

6.6. Memory Usage and Execution Times for Combining Gaussian Mixture Model and ECLAT Algorithm

A variety of data quantities were used to assess a system's performance. The execution time was 13.84 seconds, and the memory use was 217.648 MB for 100,000 items. For 200,000 items, the execution time increased to 39.09 seconds, although the memory consumption increased slightly to 239.98 MB. When scaling up to 300,000 items, the execution time increased to 41.21 seconds, and the memory use climbed to 240.02 MB. According to Table 6, memory use peaked at 249.88 MB at 400,000 items, representing an execution time of 42.34 seconds.

Table 6: Comparison of Memory Usage and Execution Time between 100,000 and 400,000 Data Items

Data Items	Memory Usage (MB)	Execution Time(s)
100000	217.648	38.93
200000	239.98	39.09
300000	240.02	41.21
400000	249.88	42.34

By comparing the two proposed methods, combining the Gaussian Mixture Model and ECLAT algorithm consumes a little more memory usage and execution time than combining the K-means and ECLAT algorithms because the Gaussian Mixture Model can cluster more items than the K-means algorithm.

7. Conclusion

This study cluster sizable datasets of UK customer descriptions using the Gaussian Mixture Model and K-means clustering. For datasets containing 100,000–400,000 elements, the optimal clustering findings indicate that 8 clusters work best for the K-means algorithm and 14 clusters are the optimal value of k for the Gaussian Mixture Model. The ECLAT method then sets a minimum support of 0.00001 to find the support items. Different clustering results are obtained by the suggested approaches with the same silhouette score of 0.59. This technique influences marketing and product strategy by identifying important consumer preferences and promoting top-selling products. When the system was tested with varying amounts of data, the memory usage varied from more than 200 MB for 100,000 items to less than 500 MB for 400,000 items, which corresponded to an execution time of less than 50 seconds. By Summarizing, the performance of ECLAT algorithm more enhance by combining Gaussian Mixture Model than K-means because Gaussian Mixture Model can more cluster the same items. Combining Gaussian Mixture Model and ECLAT algorithm can more effectively retrieve the most best selling items than combining K-means and ECALT algorithm**.**

Acknowledgement

First and foremost, I would like to sincerely thank the Marubeni Myanmar Scholarship Program. Throughout my PhD career, this program helped me improve my research every year. I also like to thank YTU's Computer Engineering and Information Technology Department for making it possible to use a powerful personal computer that was supplied by JICA from Japan to finish this research.

References

[1] Liu, Y., Liao, W. K., Choudhary, A. N., & Li, J. (2008). “Parallel Data Mining Algorithms for Association Rules and Clustering,” In Intl. Conf. on Management of Data, pp.1-25.

[2] R. Agrawal, T. Imielminski, A. Swami: “Mining Association Rules Between Sets of Items in Large Databases”. In: Proc. ACM Intern. Conf. on Management of Data, pp. 207-216, ACM Press (1993)

[3] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in Third International Conference on Knowledge Discovery and Data Mining, 1997.

[4] G. Naga Chandrika, G. Varshith, N.Bhargav Reddy & G.Gurubrahmaiah. “Customer Segmentation using K-means and Gaussian Mixture”. Journal of Engineering Science, vol 13, pp.744-750, June.2022.

[5] M. Hafidh Raditya, Indwiarti, and A. Atiqi Rohmawati, “House Prices Segmentation Using Gaussian Mixture Model-Based Clustering”, Jurnal Resti, Vol. 6 No. 5, pp. 866 – 871.

[6] N. P. Dharshinni, H. Mawengkang and M. K. M Nasution, “Mapping of medicine data with k-means and apriori combinations based on patient diagnosis”, IOP Conf. Series: Journal of Physics: Conf. Series 978 (2018) 012027.

[7]C.P. Ezenkwu, S. Ozuomba, C. Kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”, Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.10, 201

[8] A Comparative Study of Support Vector Machine and Artificial Neural Network for Option Price Prediction , **Journal of Computer and Communications,** Vol.9 No.5, May 28, 2021, Biplab Madhu, Md. Azizur Rahman, Arnab Mukherjee, Md. Zahidul Islam, Raju Roy, Lasker Ershad Ali.

[9] http://towardsdatascience.com/k-mean clustering algorithm

[10] https://www.educative.io/answers/what-is-silhouette-score

[11] https://www.deepchecks.com/glossary/gaussian-mixture-model

[12] https://quality-life.medium.com/eclat-algorithm-in-machine-learning

[13] https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-data