International Journal of Computer Sciences and Engineering
Vol.**11**, Special Issue.**1**, pp.**141-148**, November **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

An Empirical Comparison and Effect of Clustering Massive Data on Association Rule Mining

Sanjib Saha[1]

[1]Dept. of Computer Science and Engineering, National Institute of Technology, Durgapur and Dr. B. C. Roy Engineering College, Durgapur, India

Author's Mail Id: sanjib.saha@bcrec.ac.in

Abstract: This paper explores the different techniques of association rule mining (ARM) and clustering in unsupervised learning and data mining. As many works have already been done on the Apriori algorithm of ARM, but there was very limited work on the other algorithms such as Predictive Apriori, Tertius and Filtered Associator. The main problem of ARM is handling a large dataset and then scanning it repeatedly. A pre-clustering effort would reduce the dataset size for each such scan for each such cluster and thus would offer overall less time requirement. The different algorithms of ARM are executed on two different datasets such as Breast Cancer and Zoo. There is a scope for improvement in performance by applying filters and clustering techniques on ARM. The best model has been proposed as follows: (i) Use data source; (ii) Apply filters (numeric to nominal and replace missing value); (iii) Apply additional filters (attribute selection or merge two values or remove folds) or evaluation method (training set maker); (iv) Apply clustering methods (K-Means, Farthest Fast, Expectation Maximization, Hierarchical and Make Density Based); (v) Apply ARM methods (Apriori, Predictive Apriori, Tertius and Filtered Associator); (vi) View result. The different ARM algorithms are evaluated with certain metrics and compared against each other based on accuracy, lift value and execution time. However, the best rules found from each ARM algorithm are different. The paper discusses the effect of clustering on ARM and claims that clustering the data before applying ARM is better.

Keywords: Unsupervised Learning, Data Mining, Association Rule Mining, Apriori, Clustering, K-Means

1. Introduction

Data Mining is the process where data patterns of different cases are automatically classified. A variety of algorithms have been created and put into use to extract data and find knowledge patterns that could be helpful for decision assistance. Several data mining techniques are pattern association, clustering and classification.

Customer behavior in retail, banking, the healthcare system, and other industries can be studied via Association Rule Mining (ARM) [1-11]. The statements known as association rules are used to determine the correlation between the data in any dataset. "Antecedent" and "Consequent" are the two elements of the association rule. The item that was discovered in the dataset is the antecedent, and the item that was discovered beside the first is the consequent. During the process of looking for recurring patterns, association rules are formed. Identifying frequently occurring item sets and then extracting rule-based associations from such item sets are the two subproblems that make up the larger challenge of identifying association rules. Association finds all the rules X, $Y \rightarrow Z$ with minimum support and confidence. It means in retail stores if customer buys X, Y he is likely to buy Z [12, 13]. In the association rule mining research area, first effort is to improve the algorithmic performance and the second step is to reduce the output set based on constraints on the desired results.

Clustering [14] is the technique to group objects according to their similarities into one cluster and according to their dissimilarities into other clusters. That is, a high intra-cluster similarity and a low inter-cluster similarity are the goals of the objective function. Reduce the size of large data sets by clustering relevant documents for browsing, putting similar functional genes and proteins together, or grouping stocks with similar price changes.

2. Related Work and Background Method

In this paper, we have applied three types of clustering techniques on following four ARM techniques and tried to compare their performances.

2.1 Apriori Association Rule Mining

As long as those item sets exist frequently enough in the transactional database, Apriori [1] is an algorithm for identifying frequent item sets to define association rules and applying them to increasingly larger item sets. Apriori algorithm consists of candidate generation step followed by pruning step. The algorithm extended frequent subsets one item at a time and terminates when no successful extensions are found by using bottom-up approach and breadth-first search with a hash tree structure to count candidate item sets efficiently. The conditional probability, *P(Y|X)*, which is typically calculated, is confidence. Confidence demonstrates the strength of the rule, while support demonstrates the statistical relevance of the rule. The company establishes minimal support and confidence values, and all rules with more support and confidence are looked up in the dataset. It is anticipated that the lifting factor will be near 1 if *X* and *Y* are independent; but, if the ratio differs—that is, if *P(Y|X)* and *P(Y)* are different—it is anticipated that there will be a correlation between the two items: It can be observed that having *X* increases the likelihood of having *Y* if the lifting factor is greater than 1, and vice versa if the lifting factor is less than 1 [15].

Support of the association rule X →Y:

$$Support(X,Y) = P(X,Y)$$

$$= \frac{\#\{Customers\_who\_bought\_X\_and\_Y\}}{\#\{Customers\}}$$

Confidence of association rule X →Y:

$$Confidence(X \to Y) \equiv P(X \mid Y) = \frac{P(X,Y)}{P(X)}$$

$$= \frac{\#\{Customers\_who\_bought\_X\_and\_Y\}}{\#\{Customers\_who\_bought\_X\}}$$

Lift, also known as interest of association rule X →Y:

$$Lift(X \to Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(XY)}$$

2.1.1 Principles of Apriori

It is not advisable to generate a superset for an infrequent itemset.

- To get a frequent 1-itemset, first scan the dataset once.
- From length k frequent itemsets, produce length (k+1) candidate itemsets.
- Compare the candidates to the dataset.
- When no candidate or frequent set can be formed, terminate.

2.1.2 Apriori Algorithm

```
C_k: Candidate itemset of size k
$C_k = A(X)$ is the set of all one-itemsets, k = 1
while $C_k \neq \Phi$ do
scan database to determine support of all $a_y$ with $y \in C_k$
extract frequent itemsets from $C_k$ into $L_k$
generate $C_{k+1}$
$k = k + 1$
end while
```

2.1.3 Time Complexity of Apriori Algorithm

The complexity factor for Apriori algorithm is determined by:

$C = \sum_k m_k k$ where $m_k = |C_k|$

It is noted that dealing with frequent items takes up a significant amount of time. We are aware that every single item must be taken into account. Additionally, there wouldn't be items that, by themselves, are infrequent and so one has $m_2 = d(d-1)/2$. As a result, we have the lower bound for $C: C \leq m_1 + 2m_2 = d^2$

Including the data volume we find for the time complexity of Apriori:

$$T = O(d^2 n)$$

Where, d = number of items, n= number of data records.

2.1.4 Advantages of Apriori

- Makes use of huge itemset property.
- Simple to parallelize.
- Simple to develop and apply.

2.1.5 Disadvantages of Apriori

- Assumes transaction database is stored in memory.
- Needs repeated database scans.

2.2 Predictive Apriori Association Rule Mining

Support and confidence are merged into a single metric called prediction accuracy in this method [16]. The Apriori association rule is created using this level of predictive accuracy. Depending on the value of 'n' given by the user, this prediction algorithm generates the best 'n' number of association rules [17]. Predictive analytics, an area of data mining, predicts trends and behavior patterns by extracting information from data. The process of creating or selecting a model to try to most accurately anticipate the probability of a result is known as predictive modelling.

2.2.1 Advantages of Predictive Apriori

- Result will be better in Predictive Apriori than Apriori.
- By automatically resolving the balancing problem between these support and confidence criteria, the predictive Apriori algorithm maximises the probability of producing an accurate prediction for the data set.

2.2.2 Disadvantages of Predictive Apriori

- The main disadvantages of association rule algorithms in electronic learning are the use of algorithms with too many parameters for someone who is not an expert in data mining and the rules that are produced, the majority of which are uninteresting and difficult to understand.

2.3 Tertius Association Rule Mining

According to the confirmation measure, this algorithm determines the rule. It employs representation of first order logic. It enables the user to pick the most practical or understandable representation from a variety of available representations. The Tertius [18] algorithm builds rules and grades them based on how often the rule holds true in the training data, or how dependable they are. The Tertius algorithm's rules are divided into a body and a head. The literals (conditions) required for the rule to hold up make up the body. It can include a variety of conditions. If the rules are true, the event is contained in the head. Tertius begins rule learning with an empty rule-one that has an empty body and an empty head. In order to filter the rule, attribute-value pairs are attached to it as they exist in the dataset. The algorithm then counts the instances in which the rule is true or false. The Tertius rule is true positive when both body and head are true. It gives false positive when the body is true but the head is false. Tertius generally works perfectly only when the values of the attributes are categorical, otherwise it interrupts.

To find out the optimality we check the Hypothesis Considered and Hypothesis explored in the result. Using these support and confidence parameter we have found the accuracy of the rules. Let, Number of Hypothesis Considered=HC, Number of Hypothesis Explored=HE.

Therefore, Accuracy = HE / HC * 100.

2.3.1 Advantages of Tertius

- As Tertius works on the confirmation measure so it gives us the novelty and satisfaction of a rule.
- Tertius always maintain the integrity of a rule.
- It gives us the best results even for the big databases.
- Using Tertius we can find out the true positive value of a rule i.e. how much a rule is reliable.
- In each scan we can find out the accuracy measure.
- It can automatically arrange the rules based on their reliability.

2.3.2 Disadvantages of Tertius

- It takes a lot of time for implementation**.**
- In this case the memory consumption is very high.
- This algorithm cannot handle the numeric values.
- This algorithm is hard to implement.

2.4 Filtered Associator Rule Mining

An association rule mining method known as "Filtered Associator" [19] applies an arbitrary associator to data after it has been passed through an arbitrary filter. The training and test sets of data are processed by a filtered associator without altering their structure. It offers options like associator, which allows us to take into account the Tertius, Predictive Apriori, and Apriori association rules and Filtered Associator algorithm. A database of 5000 transactions with 50 unique items was used for the comparative analysis of the traditional Apriori and filtered approaches. In order to develop a frequent pattern with a 10% support count throughout this analytical procedure, we took into account 1000 transactions. The process is then repeated while gradually increasing the transaction volume. Finally, an analysis reveals that the filtering-based strategy requires 90% less time than the traditional Apriori method. As a result, this method saves about 10% of the time.

2.4.1 Algorithm of Filtered Associator

```
Initialize: $k := 1, C_1$ = all the 1-item sets;
read the database to count the support of $C_1$ to determine $L_1$.
$L_1$ := {frequent 1-item sets};
k :=2; //k represents the pass number// while ($L_{k-1} \neq$ ) do
begin
$C_k$ := generate_candidate_itemsets with the given $L_{k-1}$
prune($C_k$)
for all transactions t whose length is greater than or equal to k T do
increment the count of all candidates in $C_K$ that are contained in t;
$L_k$ := All candidates in $C_k$ with minimum support;
k := k + 1;
end
```

2.4.2 Advantages of Filtered Associator

- It saves time in comparison to all other approaches.

2.4.3 Disadvantages of Filtered Associator

- Assumes transaction database is stored in memory.
- Needs repeated database scans.

2.5 Partitioning Clustering

Assume that *n* objects in Euclidean space are present in a data set, *D*. The objects in *D* are distributed among k clusters using partitioning algorithms. In order for objects inside a cluster to be similar to one another yet different from objects in other clusters, the partitioning quality is evaluated using an objective function. A centroid-based partitioning (K-Means Clustering) [20] uses the *centroid* or *center point* of a cluster which can be defined according to the medoid or mean of the objects (or points) assigned to the cluster.

2.6 Hierarchical Clustering

In some cases, we may want to divide our data into groups at different levels, such as in a hierarchy [21], to cluster data into exclusive groups. Data objects are organized into a hierarchy or "tree" of clusters using hierarchical clustering. It is advantageous to represent data objects as a hierarchy for data summary and visualization.

2.7 Density Based Clustering

To create clusters with a spherical shape [22], partitioning and hierarchical clustering are combined. They struggle to locate clusters of arbitrary shapes, such oval and "S" clusters. We can categorize the clusters in the data space as dense regions and sparse regions in order to locate clusters of any shape. This approach works with the density-based clustering technique.

3. Material

The WEKA (Waikato Environment for Knowledge Analysis) [23] is a data mining or machine mining tool applied for data analysis and predictive modeling through a graphical user interfaces that provides easy access to this functionality. Weka supports specifically data preprocessing, clustering, classification, regression, visualization, and feature selection are included. Weka version 3.6.10 is used here. It is developed by The University of Waikato. It is a java based GUI tool applied for knowledge or data analysis and machine learning. It is open source software and it supports various format of dataset but by default it supports arff format.

Here, Breast Cancer and Zoo datasets are obtained from the UCI Machine Learning Repository [24]. Relationships of each datasets are denoted by n (numeric), c (categorical or nominal), and k (constant) and one constant should not be selected as it is class.

Table 1. Description of Datasets

Dataset	Number of Instances	Number of Attributes	Number of Classes	Relationship
Breast Cancer	286	nominal =10	2	y=1+9c+k

Zoo	101	nominal= 17,
		numeric=1
			7	y=1+16c+n+k

4. Experimental Method and Design

In this paper we have used weka as our knowledge analysis tool [25-30]. Using this tool we have prepared two different models. Using these models we extract the optimum result from the datasets that we have used. The Knowledge flow model consisting of the four association rule mining algorithm along with popular two clustering algorithm and other filtering methods from weka tools executed on Breast Cancer and Zoo datasets are given below.

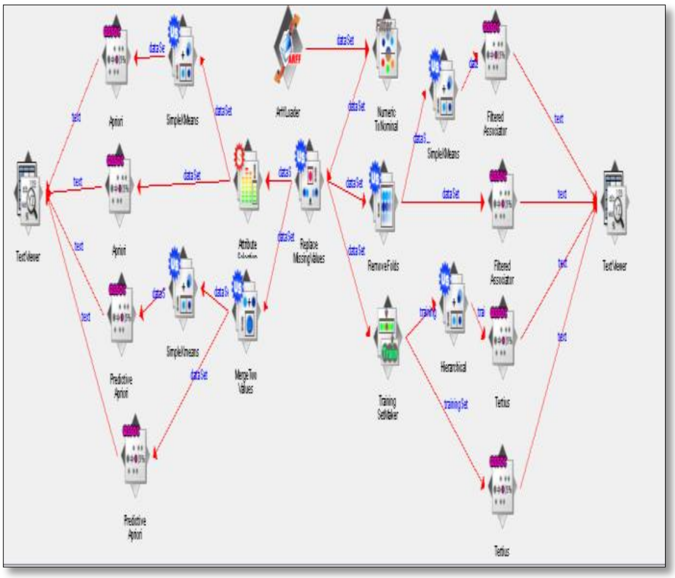


Figure 1. KnowlegdeFlow Model for Breast Cancer Dataset

In the above Figure 1 we have used the following settings-

i. Apriori-

MinLowerBound- 0.3 to 0.6 and MetricType- lift

ii. Tertius-

Confirmation Threshold- 0.475 to 0.479

Frequency Threshold- 0.5

Rests of the algorithms were executed with default settings.

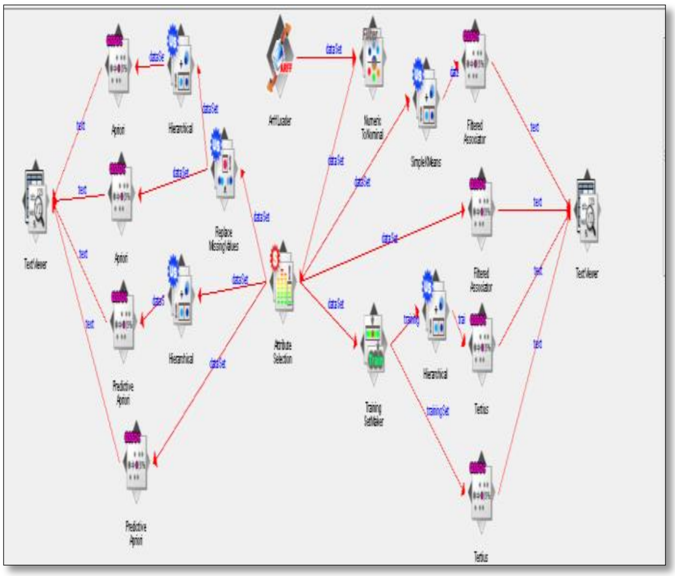


Figure 2. Knowledge Flow Model for Zoo Dataset

In the above Figure 2 we have used the following settings-

i. Apriori-

MinLowerBound- 0.3 to 0.5, MetricType- lift and MinMetric- 1.3

ii. Tertius-

Confirmation Threshold- 0.50 to 0.58

Frequency Threshold- 0.6

The components of Weka which we are used in the knowledge flow are described below.

i. Data Source
   a) ArffLoader- ArffLoader loads the dataset with .arff extension.

ii. Evaluation
   a) TrainingSetMaker- TrainingSetMaker makes a dataset into training set.

iii. Filter
   a) ReplaceMissingValues- All missing values for nominal and numeric attributes are replaced using modes and means method on the training data set.
   b) AddCluster- A new nominal attribute is introduced to indicate the cluster that the designated clustering algorithm assigns to each instance.
   c) AttributeSelection- A filter for supervised attributes that can be used to choose attributes.
   d) NumericToNominal- A filter that converts numerical characteristics into nominal.
   e) MargeTwoValues- Combines two values for a nominal property to create a single value.
   f) RemoveFolds- Fold for cross validation is specified on the training data set.

iv. Association
   a) Apriori- The minimum support is iteratively decreased in the class implementation of the Apriori-type algorithm until the necessary number of association rules are discovered with the specified minimum confidence.

b) Predictive Apriori- The class implementation of the predictive Apriori algorithm looks for the best 'n' number of rules with respect to a support-based corrected confidence value with a rising support threshold. It is possible to use the algorithm for extracting class association rules.

c) Tertius- Finds rules according to confirmation measure (Tertius-type algorithm)

d) Filtered Associator- Class for applying arbitrary associators to data that has passed through arbitrary filters. Similar to the associator, the filter's structure is solely derived from training data, and test instances will be handled by the filter without experiencing any structural changes.

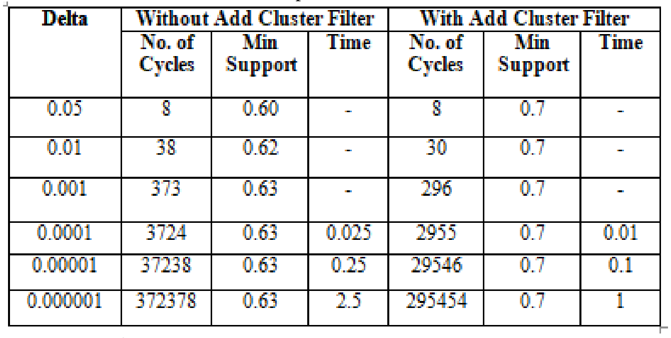
v. Visualization

a) Text Viewer- Shows the result in the text format.

5. Results and Discussion

This paper consists of a comparative analysis between the association rule mining without clustering the data and with clustering the data. The normalized time has been calculated.

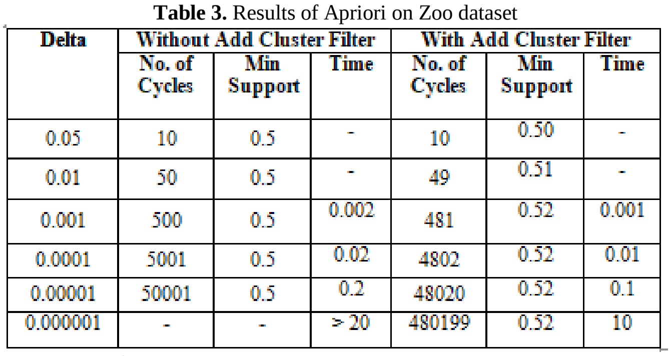
Table 2. Results of Apriori on Breast Cancer dataset



Delta	Without Add Cluster Filter			With Add Cluster Filter		
	No. of Cycles	Min Support	Time	No. of Cycles	Min Support	Time
0.05	8	0.60	-	8	0.7	-
0.01	38	0.62	-	30	0.7	-
0.001	373	0.63	-	296	0.7	-
0.0001	3724	0.63	0.025	2955	0.7	0.01
0.00001	37238	0.63	0.25	29546	0.7	0.1
0.000001	372378	0.63	2.5	295454	0.7	1

Clustering before Apriori reduces execution by 2.5 times

Table 3. Results of Apriori on Zoo dataset



Delta	Without Add Cluster Filter			With Add Cluster Filter		
	No. of Cycles	Min Support	Time	No. of Cycles	Min Support	Time
0.05	10	0.5	-	10	0.50	-
0.01	50	0.5	-	49	0.51	-
0.001	500	0.5	0.002	481	0.52	0.001
0.0001	5001	0.5	0.02	4802	0.52	0.01
0.00001	50001	0.5	0.2	48020	0.52	0.1
0.000001	-	-	> 20	480199	0.52	10

Clustering before Apriori reduces execution by 2 times

Add Cluster filter was added with the Apriori algorithm to compare the results with and without clustering the data. Breast Cancer was executed with Add Cluster filter using KMeans Cluster and Zoo dataset was executed with Add Cluster using Hierarchical Cluster on the basis of the significant rules provided by them.

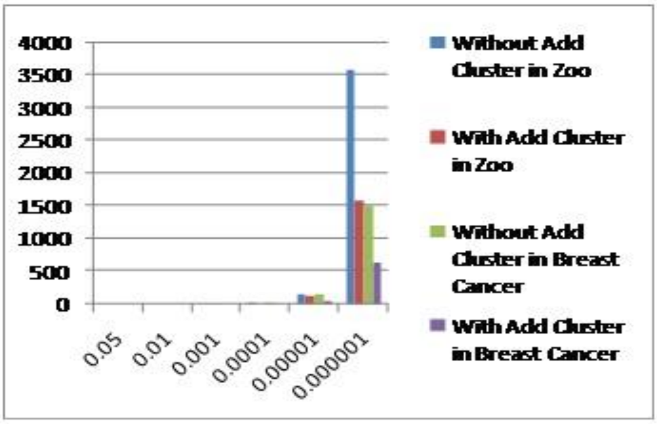


Figure 3. Performance Analysis of Apriori (Delta Vs Time)

Table 4. Results of Predictive Apriori on Breast Cancer dataset (CAR= Class Association Rules)

Feature	CAR	Class Index	No. of Clusters	No. of Rules	Time
Without Add Cluster	False	-1	-	30	1.5
With Add Cluster (K-Means)	False	-1	2	30	1

Table 5. Results of Predictive Apriori on Zoo dataset

Feature	CAR	Class Index	No. of Clusters	No. of Rules	Time
Without Add Cluster	False	-1	-	30	1.8
With Add Cluster (Hierarchical)	False	-1	7	30	1

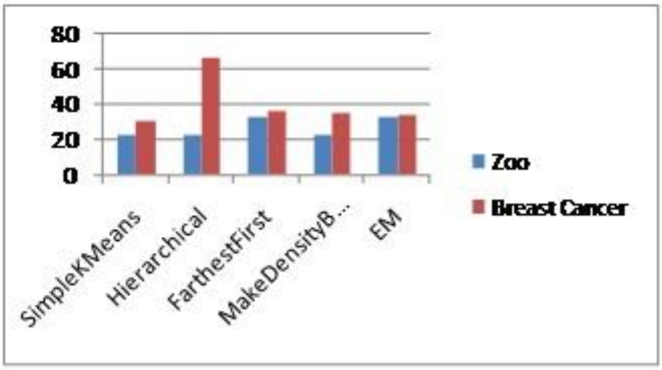


Figure 4. Performance Analysis of Predictive Apriori (Clustering Methods Vs Time)

Predictive Apriori with Add Cluster filter does not make a significant change in terms of accuracy or any other parameter and as the time change is slightly varied. We can use add cluster with the data as it does not disrupt with the normal execution of Predictive Apriori. The performance of both datasets with the different Clustering algorithms is given in the above graph. SimpleKMeans clustering gives the best result in combination with Predictive Apriori in terms of time and significant rules for Breast Cancer and Zoo dataset in comparison with other clusters.

Table 6. Results of Tertius (without clustering and with hierarchical clustering) on Breast Cancer & Zoo dataset

Dataset	CT	FT	HC	HE	NC	NR	A	T
Breast Cancer	0.475 (woc	0.5	81022	33431	-	7	41.26	2
	0.477 (wc)	0.5	110149	47434	2	6	43.06	1
Zoo	0.50 (woc	0.6	178852	72833	-	112	40.72	2
	0.50 (wc)	0.6	247124	105437	2	140	42.66	1

Confirmation Threshold(CT), Frequency Threshold(FT), Hypothesis Considered(HC), Hypothesis Explored(HE), No. of Cluster(NC), No. of Rules(NR), Accuracy %(A), Time (T)

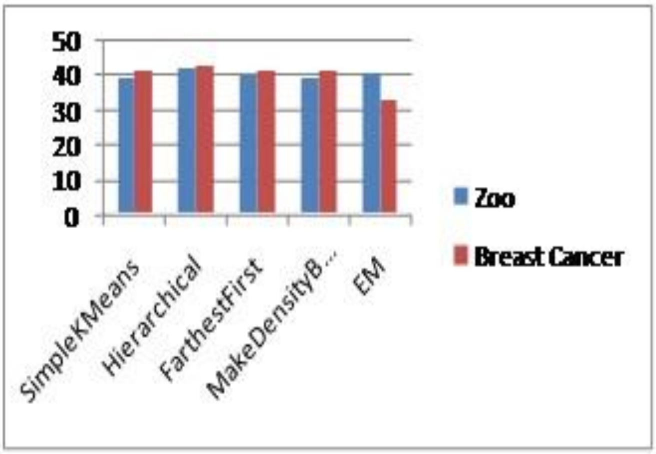


Figure 5. Performance Analysis of Tertius (Clustering Methods Vs Accuracy)

Table 7. Results of Filtered Associator on Breast Cancer dataset

Feature	Lift	Confidence	No. of Clusters (NC)	No. of Rules
Without Add Cluster	1.25	0.86	-	20
With Add Cluster (all)	1.31	0.86 to 0.95	2	20

Table 8. Results of Filtered Associator on Zoo dataset

Feature	Lift	Confidence	No. of Clusters (NC)	No. of Rules
Without Add Cluster	1.6	0.97 to 1	-	20
With Add Cluster (KMeans, MakeDensity, Hierarchical)	1.68 or 1.63	1	2 or 7	20 or 16

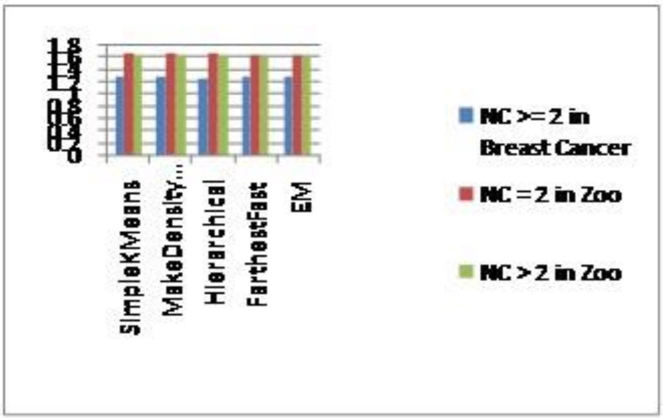


Figure 6. Performance Analysis of Filtered Associator (Clustering Methods Vs Lift)

Adding cluster to the association model significantly improves the lift value of the best rules for both dataset when no. of cluster = 2 in all cases. But increasing and decreasing the no. of clusters in Breast Cancer and Zoo dataset respectively degrade the result.

The best and common rules produced in Breast Cancer data for both the cases (without clustering and with clustering) are:

Apriori:
1. inv-nodes=0-2 213 ==> node-caps=no irradiat=no 179 conf:(0.84) < lift:(1.26)>
2. node-caps=no irradiat=no 190 ==> inv-nodes=0-2 179 conf:(0.94) < lift:(1.26)>

Tertius:
1. /* 0.492189 0.900000 0.352041 */ menopause=premeno ==> age=40-49
2. /* 0.482040 0.881720 0.352332 */ menopause=premeno ==> tumor-size=5-9 or age=40-49
3. /* 0.482040 0.881720 0.352332 */ menopause=premeno ==> inv-nodes=15-17 or age=40-49
4. /* 0.481172 0.852941 0.342391 */ menopause=premeno ==> tumor-size=35-39 or age=40-49

Filtered Associator:
1. inv-nodes=0-2 21 ==> node-caps=no irradiat=no 18 conf:(0.86) < lift:(1.31)> lev:(0.15) [4] conv:(1.81)
2. node-caps=no irradiat=no 19 ==> inv-nodes=0-2 18 conf:(0.95) < lift:(1.31)> lev:(0.15) [4] conv:(2.62)

The best and common rules produced in Zoo data for both the cases (without clustering and with clustering) are:

Apriori:
1. hair=false 58 ==> milk=false 56 conf:(0.97) < lift:(1.63)>
2. milk=false 60 ==> hair=false 56 conf:(0.93) < lift:(1.63)>

Tertius:
1. /* 0.602707 1.000000 0.403226 */ feathers=false and fins=false ==> legs=4 or animal=clam
2. /* 0.602707 1.000000 0.403226 */ feathers=false and fins=false ==> legs=4 or animal=crayfish
3. /* 0.602707 1.000000 0.403226 */ feathers=false and fins=false ==> legs=4 or animal=flea
4. /* 0.602707 1.000000 0.403226 */ feathers=false and

fins=false ==> legs=4 or animal=fruitbat

Filtered Associator:

1. toothed=true 61 ==> feathers=false backbone=true 61 conf:(1) < lift:(1.68)> lev:(0.23) [22] conv:(22.95)
2. feathers=false backbone=true 63 ==> toothed=true 61 conf:(0.97) < lift:(1.68)> lev:(0.23) [22] conv:(8.32)

6. Conclusion

In this paper, after comparing the various results and graphical analysis of all the association rule mining algorithms such as Apriori, Predictive Apriori, Tertius and Filtered Associator, it can conclude that after clustering the data and then applying association, we are getting an improved result than applying only association without clustering the data. When association rule mining is executed after clustering the data then for considerably small values of delta in the Apriori property, the time taken to generate the rules is significantly less than the time taken by association alone in case of Apriori. On the other hand, Tertius has shown a noticeable level of increase in the accuracy of the rules when association rule mining was done after clustering of the data. Filtered Associator ensures that there is an increase in the lift value of the rules after clustering the data. Hence improves interestingness of the best rules. When these association rule mining algorithms were compared against each other, it is found that the metric type of these algorithms to determine the best rules was different in each case. Thus, Execution time is the only metric to draw a comparison between these algorithms. Tertius always gives only the best possible rules that can be determined from the dataset in every case. Those rules have all the best possible combinations of the attributes within them and they are different from the monotonic rules. Apriori takes the minimum execution time to give the best rules in case of all the datasets in comparison with the other algorithms in default settings. However, Filtered Associator using Apriori gives better lift value than only Apriori.

Conflict of Interest

There is no conflict of interest for this article.

Funding Source

There is no funding source for this article.

Authors’ Contributions

The single author of this article is the only contributor.

Acknowledgements

The author is thankful to his parents, wife, and daughter for their support.

References

[1] Agarwal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. of the 20th VLDB Conference. Vol.**487, 1994.**

[2] Agrawal, Rakesh, Tomasz Imielinski, and Arun Swami. "Database mining: A performance perspective." IEEE transactions on knowledge and data engineering 5.6: pp.**914-925, 1993.**

[3] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." Proceedings of the 1993 ACM SIGMOD international conference on Management of data. **1993.**

[4] Kaushik, Minakshi, et al. "A systematic assessment of numerical association rule mining methods." SN Computer Science 2.5: **348, 2021.**

[5] Ünvan, Yüksel Akay. "Market basket analysis with association rules." Communications in Statistics-Theory and Methods 50.7: pp.**1615-1628, 2021.**

[6] Kaushik, Minakshi, et al. "On the potential of numerical association rule mining." Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, Proceedings 7. Springer Singapore, 2020, November pp.**25–27, 2020.**

[7] Liu, Bing, Yiming Ma, and Ching Kian Wong. "Improving an association rule based classifier." Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, Proceedings 4. Springer Berlin Heidelberg, 2000, September pp.**13–16, 2000.**

[8] Altaf, Wasif, Muhammad Shahbaz, and Aziz Guergachi. "Applications of association rule mining in health informatics: a survey." Artificial Intelligence Review 47: pp.**313-340, 2017.**

[9] Kaur, Manpreet, and Shivani Kang. "Market Basket Analysis: Identify the changing trends of market data using association rule mining." Procedia computer science 85: pp.**78-85, 2016.**

[10] Feng, Feng, et al. "Soft set based association rule mining." Knowledge-Based Systems 111: pp.**268-282, 2016.**

[11] Chiclana, Francisco, et al. "ARM–AMO: An efficient association rule mining algorithm based on animal migration optimization." Knowledge-Based Systems 154: pp.**68-80, 2018.**

[12] Ganda, Ritu. "Knowledge discovery from database using an integration of clustering and association rule mining." International Journal of Advanced Research in Computer Science and Software Engineering 3.9: pp.**13-18, 2013.**

[13] Shweta, Ms, and Dr Kanwal Garg. "Mining efficient association rules through apriori algorithm using attributes and comparative analysis of various association rule algorithms." International Journal of Advanced Research in Computer Science and Software Engineering 3.6: pp.**306-312, 2013.**

[14] Tan, Steinbach, and Kumar, "Cluster Analysis: Basic Concepts and Algorithms," Introduction to Data Mining, 2006, Addison-Wesley.

[15] Yılmaz, Nergis, and Gülfem Işıklar Alptekin. "The Effect of Clustering in the Apriori Data Mining Algorithm: A Case Study." Proceedings of the World Congress on Engineering. Vol.**3. 2013.**

[16] Scheffer, Tobias. "Finding association rules that trade support optimally against confidence." Intelligent Data Analysis 9.4: pp.**381-395, 2005.**

[17] Aher, Sunita B., and L. M. R. J. Lobo. "A comparative study of association rule algorithms for course recommender system in e-learning." International Journal of Computer Applications 39.1: pp.**48-52, 2012.**

[18] Flach, Peter A., and Nicolas Lachiche. "Confirmation-guided discovery of first-order rules with Tertius." Machine learning 42.1-2 (2001): 61.

[19] Bathla, Himani, and K. Kathuria. "Apriori algorithm and filtered associator in association rule mining." International Journal of Computer Science and Mobile Computing 4.6 (2015): 299-306.

[20] MacQueen, J. "Classification and analysis of multivariate observations." 5th Berkeley Symp. Math. Statist. Probability. Los Angeles LA USA: University of California, 1967.

[21] Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.1 (2012): 86-97.

[22] Kriegel, Hans-Peter, et al. "Density-based clustering." Wiley interdisciplinary reviews: data mining and knowledge discovery 1.3 (2011): 231-240.

[23] WEKA3 tool for machine learning and knowledge analysis. Online available at http://www.cs.waikato.ac.nz/~ml/weka/

[24] Blake, C. and Merz, C. J. "UCI repository of machine learning datasets." University of California, Irvine, Dept. of Information and Computer Sciences.(http://www.cs.waikato.ac.nz/~ml/weka/)

[25] Asadi, Sh, Seyed Jafari, and Z. Shokrollahi. "Developing a course recommender by combining clustering and fuzzy association rules." Journal of AI and Data mining 7.2: pp.**249-262, 2019.**

[26] Datta, R. P., and Sanjib Saha. "Applying rule-based classification techniques to medical databases: an empirical study." International Journal of Business Intelligence and Systems Engineering 1.1: pp.**32-48, 2016.**

[27] Saha, Sanjib, and Debashis Nandi. "Data Classification based on Decision Tree, Rule Generation, Bayes and Statistical Methods: An Empirical Comparison." Int. J. Comput. Appl 129.7: pp.**36-41, 2015.**

[28] Das, Subhankar, and Sanjib Saha. "Data mining and soft computing using support vector machine: A survey." International Journal of Computer Applications 77.14, **2013.**

[29] Saha, Sanjib. "Non-rigid Registration of De-noised Ultrasound Breast Tumors in Image Guided Breast-Conserving Surgery." Intelligent Systems and Human Machine Collaboration. Springer, Singapore, pp.**191-206, 2023.**

[30] Saha, Sanjib, et al. "ADU-Net: An Attention Dense U-Net based deep supervised DNN for automated lesion segmentation of COVID-19 from chest CT images." Biomedical Signal Processing and Control 85: 104974, **2023.**

AUTHORS PROFILE

Sanjib Saha earned his Bachelor of Engineering and Master of Technology from Burdwan University and Jadavpur University respectively. He is pursuing PhD in Computer Science and Engineering at National Institute of Technology, Durgapur, India. He is working as an Assistant Professor in Department of Computer Science and Engineering at Dr. B. C. Roy Engineering College, Durgapur. He has published research papers in SCI and Scopus journals including conferences. His main research work focuses on Machine Learning, Deep Learning, and Medical Image Classification, Segmentation & Registration.