An Enhanced Novel Clustering Technique for Non-Associative Item Set (CNAIS) of Transaction Data Sets

M. Vinayababu[1*] Dr.M.Sreedevi[2]

[1]Research Scholar**,** Department of Computer Science, Sri Venkateswara University, Tirupati, India.

[2]Associate Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, India.

Abstract: Generally Clustering and Associative rule mining are important sub-fields of data mining which are applied to various fields like marketing (e.g., customer segmentation), image identification and analysis, bioinformatics, text mining, document classification, indexing, Health care analysis, financial fields etc. Every feature of data set if analysed will provide some kind of assistance to decision makers in their fields which will bring profits/ show new way of marketing or getting customers to attract to product. So in this research paper we have used CNAIS (Non-Associative Items Clustering) technique which also uses items of highest threshold value (not same as in ARM) which may not found in associative mined items if clustered. Generally after laborious process of Association rule mining, Non-associated items may not be considered and analysed, but if we can do useful analysis on those items which will allow us to consider full set of transaction items we can have maximum number of customers can be considered for benefit of stake holders in the field which is in consideration of problem under study.

Keywords: Clustering Algorithm; Association Rule mining; Artificial Intelligence; CNAIS method.

I. Introduction

Due to emergence of latest technologies in Information sector such as AI, Cloud computing data mining are playing key role in their advancements. As we know that Data mining is applied on large repositories or databases or data ware houses for extracting interesting useful facts which may be significant to concept understudy , brings out hidden features and potentially useful in information gathering or unearthing patterns . One of the key process in KDD is data mining which uses different algorithms to extract hidden knowledge. Efficient algorithms are applied to data mining processes like generalization and characterization of patterns, classifying and clustering data based on features, association, evolution of new patterns, pattern matching, data visualization and meta-rule guided mining.

Mining techniques depending upon its outcomes which focuses only on certain features of data in the database as they selects only certain features which are related to process from the integrated data resources with varieties of data and transform them into a form suitable for mining task. Different implementations of mining techniques will run on data sources which maybe huge in volume, for extracting different knowledge out comes suitable for various analysis and decision making. Those knowledge outcomes are evaluated and visualized in different ways suitable for that domain in consideration like raw data in different views, tabular form, decision tree formats, graphs, rules, charts, data cubes or multi-dimensional graphics. [1]

Data mining briefly classified as Descriptive and Prescriptive views, where descriptive mining summarizes or characterizes huge data based on general properties in data repository and inferencing and predicting data based on historical data is Prescriptive mining. Various techniques like association, clustering, classification of data items, investigating outliers, regression and trending analytics, machine learning methods are part of either of the data mining types. Mining huge volumes of data is not a simple task but throws many challenges across broad categories like Finding Missing Values, apt Feature Selection, focus on Outlier Detection, methods for Cluster Analysing of high dimensional data, Identifying Imbalanced classes in classification, providing Privacy to data, how to extract patterns from complex/distributed data, leaving data unused in data sets due to algorithm logic etc. Among the various techniques of data mining Association rule mining, one of the most frequent used techniques of data mining, was first introduced in [Agrawal et al. 1993]. ARM finds out useful correlations in

data items, frequent patterns in datasets, association among items or casual structures involved in the transaction databases or data repositories [2].

Partitioning and segmentation of data among large data sets are extracted by technique called clustering which groups data into multiple subsets termed as clusters. Based on features those objects which are nearer to each other are placed into similar groups or clusters. Similar to classification, clustering also agglomerates the similar data objects with class labels are unknown as it is an unsupervised learning technique which differs from classification in this aspect which is supervised learning. Various domains like image segmentation, pattern recognition, statistics etc., finds usage of Cluster analysis which are further classified into hierarchical, vertical clustering etc.[3][20]

II Related Works

In paper [4] by Brain Lenty et.al has developed an automated system to compute a clustering of data set with two-attribute space in large databases and how association rule mining technology can be applied to the clustering problem. In order to reduce number of passes authors have proposed a specialized mining algorithm that only makes one pass through the data set for a given partition of the input attributes. Computation of support or confidence thresholds are changed without requiring a new pass through the data. A new geometric algorithm for locating clusters in a two-dimensional grid was introduced. In this algorithm not only uses rules corresponding to specific attribute with equalities but also association rules with attributes with values inequalities are considered as a rules.

In [5][6][7] authors have provide some of the modifications to Apriori algorithm to improve efficiency by reducing database scans or reducing search space as Direct Hashing & Pruning (DHP) algorithm. Datasets are analysed by DHP which is an effective hash based algorithm for candidate item set generation. By reducing size of candidate-2 item set this algorithm reduces computational time. In [6][7] authors has modified DHP with H-Bit Array Hashing and Perfect Hashing and Pruning [PHP] respectively overcome drawbacks in DHP algorithm.

In [8] authors have proposed an improved algorithm in association rules based on Itemset Matrix(ISM) and Cluster Matrix(CM) , in this technique just one scan of database will produce frequent k-itemsets in minimum time with less complexity. Even When database are updated only by scanning changed records we can update frequent itemsets without scanning entire database thus reduces time.

Authors R. Agrawal and R. Srikant [9] has proposed GSP (Generalized Sequential pattern) algorithm efficient than AprioriAll which incorporate time constraints, sliding time windows, and taxonomies in sequential patterns. In transactions database when number of transactions increases GSP scales linearly with set of data-sequences and number of items per transaction and shows better performances than others.

Liu et.al in [10] has proposed an Apriori-based algorithm, named MSapriori which generalized association rule mining algorithm by proposing multiple minimum threshold numerical entity instead of assuming a single minimum support threshold for all items. In this technique allow users to specify multiple minimum supports to reflect the natures of the items they have purchased or used, and developed to mine all frequent itemsets. By using MIS-tree even though multiple supports counts are used, it avoids multiple scans of databases by storing key information about frequent patterns.

In [11], an improvement to projected clustering algorithm (DOC), a technique known as Frequent-Pattern-based Clustering (FPC) is proposed which applies the branch and bound paradigm to efficiently discover the projected clusters. It uses searching mechanism based on mining frequent itemsets for finding projected cluster and along with FP-Growth algorithm. FPC algorithm among the various medoids 'p' finds the best projected cluster. FPC is also extended to CFPC, a technique that gives multiple clusters concurrently at a single instance of the iterative process.

Level wise k-dimensional exploring algorithm CLIQUE proposed in [13] is one of the foremost projected clustering algorithms finds projected clusters of k^{th} dimensionality after evaluation k-1^{th} dimensionality have been discovered. Modified CLIQUE, a PROCLUS given in [12] is a medoid-based projected clustering technique which is efficient in terms of scalability of CLIQUE as it selectively takes a good number of candidate medoids and finds out nearest clusters around them. Subspaces of clusters are built around points near the medoids. In case of large clusters even though PROCLUS is fast may not be effective when considering clusters

of large variance (no medoids are chosen from smaller clusters) in that case they are split into small and taken as outliers.

In [14] application Customer Relationship Management solutions have used several technologies to understand more accurately the behaviour and requirements of customers , by engaging and making them satisfied with shopping experience which in turns increases retailers profitability. By applying multi-clustered approach of items linked with user profiles are used in recommendation systems rather than using a general approach to customer, depends on most the items customer has purchased.

In above all research works frequent pattern mining and clustering is done with transactions sets, and the application of Associative analysis to Market basket analysis is computationally expensive in discovering patterns from large set of transactional data sets. Sometimes resultant patterns extracted may be not useful as they may be extracted by chance. Large set of data which are left unused has not considered and the attributes in non-associated items sets are not used even though they are part of final resultset due to frequent patterns or others. So in this paper we have focused on applicability of non-associated items and clustering these items which would benefit business they might have left over important rows in dataset by chance.

III. ASSOCIATION RULE MINING

A patterns, also known as itemsets that appear frequently in dataset is known as frequent patterns and extracting such patterns will provides insights into associations among data items, correlations and other relationships in data sets. For example frequent pattern or a subsequence, such as buying first a PC, then a UPS, and then a wireless mouse, screen guard, if it occurs frequently in a shopping history database of customers, then it is a (frequent) sequential pattern. Today's business world is spreading 360 degrees which leads to massive data generation continuously, it is collected and stored. Many business analytics focus on mining frequent patterns in data sets stored above that will help them in making decisions effectively in terms of product-catalog design, ways of cross-marketing, and customer shopping behaviour analysis while shopping. Market basket analysis is one such an example of kind frequent Itemset mining. [15].The analysing of huge volumes of data to uncover intriguing linkages and relationships is coined as association rule mining or problem mining frequent itemsets. By using this rule we can find number of times an item set appears in a transaction. This enables stakeholders to determine connections between the products and individuals commonly purchase such products.

In general, association rule mining on dataset can be done in two steps

1. Scan datasets to find all frequent itemsets: By selecting an appropriate predefined minimum support count(min_sup_count) itemsets occurring frequently satisfying min_sup_count is extracted.

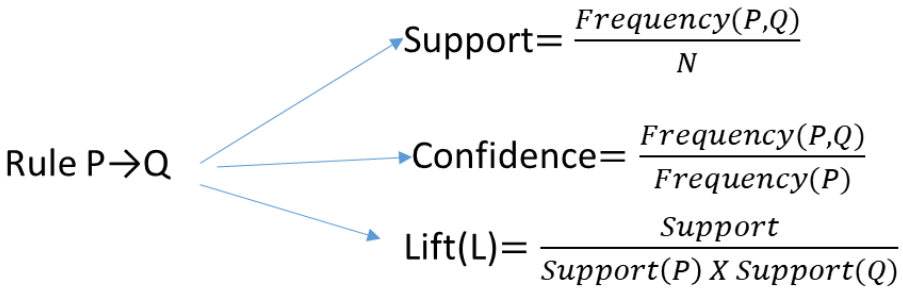
2. Generate association rules from the frequent itemsets: Associated item rules are generated-from itemsets in step-1 which all should satisfy minimum support and minimum confidence values.

The Association Rule Mining uses following parameters such as

Minimum Support Count, (min_sup_count) should be selected appropriately nearer to 60% of total transactions will determine the frequency of occurrence of an Itemset in the number of transactions that comprises the Itemset.

The Confidence(C) of an association rule is a percentage value that shows how frequently the rule head occurs among all the groups containing the rule body. $P \Rightarrow Q$ can be easily derived from the support counts of P and $P \cup Q$. After finding support counts of P, Q, and $P \cup Q$, directly corresponding association rules $P \Rightarrow Q$ and $Q \Rightarrow P$ can be checked whether they are strong.

The Lift value determines how important the rule is. By applying rule filters, appropriate lift range can be defined. It is given as the ratio of the confidence (C) of the rule ($P \Rightarrow Q$) and the expected confidence (EC) of the rule($P \Rightarrow Q$).
Lift (L) = C/EC



For example consider following Transaction DataSet

Transection ID	Item Sets
1	Biryani, Pizza, Cooldrink, fish fry, Ice-Cream
2	Gulab jam, Biryani, Tandoori, Halwa, Sandwich, Mutton fry, Lemon soda
3	Pizza, Cooldrink, Biryani, Veg-rice, Samosa, Ice-Cream
4	Biryani, Gulab jam, Halwa
5	Veg-Rice, Lemon soda, Cooldrink, Ice-Cream, Pizza, Biryani
6	Sandwich, Chicken Fingers, Mutton fry, Cool-Drink

Table:1 Associated itemsets

Let's first look at the fundamental definitions before we define the rule.

Support Count () – Frequency of occurrence of an item set.

Here ({Pizza, Ice-Cream, Cool drink }) =2

Frequent Itemset – An Itemset whose support is greater than or equal to the min_sup_count threshold.

Association Rule – An implication expression of form P →Q, where P and Q are any 2 item sets.

Example: {Pizza} => {Ice-Cream, Cool drink}

Rule Evaluation Metrics:

(i) Support(s) –The number of transactions containing elements from the "P" and "Q" portions of the rule is applied when its percentage of all transactions is calculated. The elements usually occur together as a percentage of all transactions.

(ii) Support = (P+Q) total – The percentage of transactions that contain both P and Q is how it is usually understood.

(iii) Confidence(c) – The ratio of the number of transactions, which includes all of the entries in column B as well as the number of transactions that adds all of item "A's" transactions to the number of transactions that contains everything in A.

(iv) Conf(P=>Q) = Supp (P,Q)/ Supp(P) – It gauges the frequency with which each item in Y appears in transactions that include stuff also in Lift(L) – Assuming that the item sets P and Y are independent of one another, the lift of the rule

(v) P=>Q is the confidence of the rule divided by the predicted confidence. The confidence divided by the frequency of "Q" yields the predicted confidence.

(vi) (vi) Lift(P=>Q) = Conf(P=>Q) /Supp(Q) – Greater lift values suggest stronger associations. A lift value close to 1 implies that P and Q almost always appear together as predicted. A lift value greater than 1 show that they appear together more frequently than expected

Example From the above data,

{Pizza} => {Ice-Cream, Cool drink}

Support (S)= **σ(P+Q) ÷ total**

= 3/6

= 0.5

Confidence (C)**= Supp(PUQ) ÷ Supp(P)**

=3/3

=1

Lift = Supp ÷ Supp(P) * Supp(Q)

= 0.5/(3*3)

= 0.5/9

= 0.05

When evaluating datasets, the Association rule are helpful. In Restaurants, barcode scanners are used to get the data. These databases contain a high number of transaction records that list each item a consumer has purchased in a single transaction. As a result, the manager might determine whether specific product categories are frequently bought together and utilize this information to modify store layouts, cross-sell opportunities, and promotions depending on statistic. [15]

IV TECHNIQUES IN CLUSTERING ALGORITHMS

Clustering is similar to classification, but for given a set of data objects grouping is to be done where the class label of each object is not known. Grouping in c*lustering results in creating classes or cluster of objects*, so that objects within a cluster have high similar characteristics in comparison to other, but are very dissimilar to objects in other clusters. Based on attribute values dissimilarities are considered for describing the objects. Often, distance measures, centroids etc., are used. Various fields like data mining, statistics, biology, and machine learning has its applications to clustering [16][17][18].

The clustering technique involves following process such as

a) Feature extraction and selection: extract and select the most representative features from the original data set;
b) Clustering algorithm design: design the clustering algorithm according to the characteristics of the problem;
c) Evaluation: evaluate the clustering result to verify the validity of algorithm by using practical available dataset.

There are different algorithms but they are considered based on the reasons such as choosing considering these algorithms are Flexibility for different data sets, Handling of high dimensionality, complexity and its applicability in different areas.

Accordingly, the based on mechanism of processing datasets clustering algorithms can be broadly classified as follows: [16][17][18].

a) Partitioning-based: These clustering algorithms are simple to implement and clusters are easily determined in effective manner. Initially groups are specified and finally reallocated and combined to form clusters. In other words, the algorithms like K-Means, K-Medoids, PAM, CLARA, CLARANS and FCM etc., scan data and divide data objects into a number of partitions where each partition considered as a cluster.

b) Hierarchical-based: Also known as Agglomerative or divisive partitioning. By considering proximities in attributes of Data set clusters are organized in either top-to-down in hierarchical manner. Proximities are obtained by the intermediate nodes. A dendogram is formed as algorithm scans the datasets, with leaf nodes representing individual data. The initial cluster gradually expands into hierarchical tree to form several clusters as new nodes are added. The approach in Hierarchical clustering methods are either bottom up or top-down. Examples are BIRCH, CURE, ROCK and Chameleon are some of the algorithms of this technique.

c) Density-based: In this method of building clusters regions of density, connectivity and boundary are considered when processing data objects. Cluster is a connected dense component and grows in any direction depends on density. It is shown as point-nearest to neighbours. Clusters of variable shapes are created in these density-based algorithms like DBSCAN, OPTICS, DBCLASD and DENCLUE.

d) Grid-based: Database on which clustering is to be done divides the space of the data objects into grids. Grid based approach is efficient in terms of time complexity, because it scans through the dataset once to compute the statistical values for the grids. This makes grid based algorithms independent of size of dataset .Wave-Cluster and STING are typical examples of this category.

e) Model-based: These types of clustering technique are robust techniques which optimizes the processing choosing fit between the given data and some (predefined) mathematical model by considering number outliers. The two major approaches that are considered in the model-based methods are statistical and neural network. MCLUST technique is probably the best-known model-based in terms of complexity and clusters fitting as per evaluation.

Among various clustering techniques here we discuss about basic technique K-means algorithm which is simple, efficient and most used unsupervised clustering algorithm to cluster the data. It is one the partition based method. The algorithm is implemented in two stages [19]

a) Firstly, the data is divided into 'k' number of clusters with assumed 'k' value in advance. Among the given set of data take 'k' number of points and assume it as a centroid for that respective cluster.
b) Secondly, calculate the distance between point and centroid and assign the point to the
Cluster, which has the least distance which brings the point closest centroid. This method reduces the number of iterations and change of locations of points in clusters.

Briefly the procedure is given as
1. Initially take value 'k' as a count of clusters.
2. Initialize the vectors to the these 'k' clusters taken in step-1
3 For each new vector:
3.1 Calculate the distance between the new vector and every centroid.
3.2 Compute the closest centroid and add the new vector to that respective cluster.

V. Proposed CNIAS Algorithm

In studying above Association and Clustering techniques , we propose Clustering Technique of Non-Association rule Items Sets (CNIAS) over Transaction item sets which considers non-association rule item sets to cluster them along with associated rule based sets or independently not to lose the data items or rows in database as unused after filtered in either ARM or clustering after long computation.

A. CNAIS Technique

Let $I=\{i_1,i_2,i_3\ldots i_n\}$ be the set of Items in Transaction Database T_d containing set $T\{t_1,t_2,t_3\ldots t_n\}$ of transactions of customers/stakeholders $\{c_1,c_2,c_3\ldots c_n\}$ in the context.

T_d is a database of rows where each row has 'n' number of features/attributes which store item values.

$S_A = \{ i_{a1},i_{a2},\ldots i_{an}\}$ be the set of Associated item sets which is a subset of I.

$(S_A \subset I)$

$S_{NA} = \{ i_{na1},i_{na2},\ldots i_{nan}\}$ be the set of Non- Associated item sets which is a subset of I.

($S_{NA} \subset I$)

Association rule is defined a A_r form A : $i_{a1} \rightarrow i_{a2}$ where $i_{a1}, i_{a2} \in I$ and appears more than once in rows of T_d .

Support(s) and Confidence (c) are the two measures of Associative rule

$I_{a1} => I_{a2}$, such that

Support quantifies how many times $[I_{a1}, I_{a2}]$ occurs in some of the rows as the fraction of total tuples.

ie.. S_c= |frequency ($[I_{a1}, I_{a2}]$)) /rc| , rc is total tuples in T_d

Confidence quantifies how I_{a1}, I_{a2} appears together as a fraction of number of rows in which I_{a1} occurs . ie. S_c $(I_{a1}, I_{a2})/S_c(I_{a1})$

The proposed algorithm Clustering Technique of Non-Association rule Items Sets over Transaction item sets (CNAIS) works in three phases

a) Extracting Non-Associative Rule Sets (S_{Na}) from T_d transaction dataset.

b) Computation of Threshold value and applying Clustering technique over items of both (S_A and S_{NA}) to form clusters which will enable us to group Transactions T_i in T_d according to T-value

c) Based on different T-Values we can prioritize Transactions T_i in T_d

In Phase-1 we use function ENAIS(Td,SA,SNA)

Algorithm : ENAI(T_d,S_A,S_{NA})

Inputs : Transaction Database T_d containing set T{t_1,t_2,t_3…t_n}

Returns : S_A,S_{NA}

1. Begin
2. For each Ti in Td do
   2.1 Count occurrences of I_k
   2.2 Find SupportCount (Sc)
   2.3 S_A =[set of items >= Sc]
   2.4 S_{NA}=[set of items<Sc]
   End of step-2
3. Repeat until count_itemset(S_A) = Prev_Count each Item in S_A do
   (Loop until no more items satisfy)
   3.1 Prev_count= count_itemset(S_A)
   (Item set count before generating next new Item-sets)
   3.2 Calculate SupportCount(Sc)
   3.3 get Subset Sk that satisfy Supportcount(S_c)
   3.4 combine items in S_A=S_k U S_{Ak-1}
   (for generation itemsets of size k that satisfy Supportcount (Sc))
4. S_{NA}=T_d - S_A
5. return S_A,S_{NA}

In Phase-2 we use function GenerateCluster(S_A,S_{NA}) to create clusters which are less in outliers and perfect.

Algorithm :GenerateCluster(,S_A,S_{NA})
Inputs : SA = { ia1,ia2,….ian} be the set of Associated item sets which is a subset of I.

SNA = { ina1,ina2,....inan} be the set of Non- Associated item sets which is a subset of I.
Returns : Clusters $K_1..K_m$(contains S_A,S_{NA})

1. Begin
2. Select attribute A_{it} which will be used as T-Value(T_v) for selection of Item from S_A,S_{NA}
3. Arrange in ascending order of A_{it} in S_A and find Count C_n
4. Find Median of values M_{SA}
5. Repeat steps 3 & 4 for S_{NA} and find M_{SNA}
6. $T_v = M_{SA} + M_{SNA}/2$
7. Choose T_V as threshold value for creating clusters.
8. Initialize a points randomly $k_{1...}k_m$ … as number of clusters required from A_{it} of S_A,S_{NA} such that the value is nearest as that of support count (+ or - Sc %) to T_V (threshold values) from S_A and S_{NA} data points as the medoids.
9. For each points chosen from above step-8, create from n objects in S_A and S_{NA} datasets, assign them to k_m clusters such that each k_i object is assigned to one and only one cluster. Hence, it becomes an initial medoid for each cluster.
10. For all remaining non-medoids in each k_m , compute the Cost (distance as computed via Manhattan method) from initial medoid.
11. In each k_i^{th} cluster each medoid is compared with that of the k_{i+1}^{th} cluster and minimum distance (values related to SC% of threshold value Tv) is selected and clusters are formed with S_A and S_{NA}
12. Compute the total cost of min medoids in k_i^{th} i.e. it is the total sum of all the non-medoid objects distance from its cluster medoids and assign it to D_{ki}. Similarly find for k_{i+1}^{th} cluster and assign to Dk_{i+1} ie $D_{ki} = \Sigma(\min(ki))$ and $Dk_{i+1} = \Sigma(\min(k_{i+1}))$

13: Compute $S_1=|D_k-D_{k+1}|$

14. Repeat the steps for k_i th and k_{i+2}^{th} cluster step-11 to step-13 and find S_2
15. Compute Z=S1-S2
16. if (z<0) then

    Swap initial medoids to next Random medoid and Repeat steps 8 to 15 until clusters has no change

    Else

    Clusters k_i, k_{i+1}, k_{i+2} are perfect.

Return $k_i..k_m$ clusters.
end .

In Phase-3 we can use Phase-2 algorithm, to generate clusters K1…Km where 1…m are priority based clusters are created based on threshold value.

Algorithm: PriorityClusterGenerate(k1..Km)
Input: Clusters K1...Km formed using S_A and S_{NA} based on T_V which depends on A_{it}
Output: Clusters Graphs with Non-Associated items Selected and Priority based Selection.

1 Begin
2. For each Pair(K_i,K_{i+1}) do

   A.compute $x= \Sigma|k_i med + k_{i+1} med|$ and $y=k_{i+2} med$

   b) Plot scatted graph G_i

3. G_i give S=Non-Associated Items Selected for SA and SNA satisfying Threshold value T_V

   ; Prioritize Clusters points based on T_v

4. Plot graph using clusters Ki/Ki+1 Ki/K+2 with threshold Value Tv clusters are formed with different points with T_v+ Sc% of A_{it}

End.

VI Results

Consider the following Dataset select with limited rows and attributes to demonstrate our example.

Step 1: First, initialize: select the k random points from n number of customer bills.

Step 2: Choose the D1 point as compared to the min threshold value. In this D1 point we take the x value as above threshold value and y value as below threshold value from data set. And same as D2, D3.

Cus-Id	Associative Price(Ait)	Non-Associative Price(Ait)	D1 (4800, 3750)	D2 (7100, 4000)	D3 (6750, 3600)
1	7200	4500	3650	600	1850
2	7100	4000	3250	0	1450
3	4450	5300	1900	4650	4000
4	4800	3750	0	3250	2100
5	2970	6120	4200	6950	6300
6	6700	3600	2100	1450	0

Table:2 Reduced data set from transaction set chosen with Attribute (Ait) for threshold value. Red VALUES are selected as Non-Associated values from clusters formed. Here $T_v = 4812$

Step 3: For all objects we compute the Cost (distance as computed via Manhattan method) from all medoids.

Manhattan Distance b/w P and Q = |x1-x2| + |y1+y2|

Here, D1(4800,3750) and P (7100,4000)

- |x1-x2| + |y1+y2|
  = |4800-7800|+|3750-4000|
  =3000+250
  =3250

And D1(4800,3750) and P (4450,5300)

- |x1-x2| + |y1+y2|
  = |4800-4450|+|3750-5300|
  =350+1550
  =1900

Similarly, Find D2 Values
Here, D2(7100,4000) and P (7200,4500)

- |x1-x2| + |y1+y2|
  = |7100-7700|+|4000-4500|
  =600+500
  =1100

And D2(7100,4000) and P (4450,5300)

- |x1-x2| + |y1+y2|
  = |7100-4450|+|4000-5300|
  =2650+1300
  =3950

Similarly, Find D3 Values

D3(6750,3600) and P (7200,4500)

- |x1-x2|+|y1+y2|
  =|6750-7700|+|3600-4500|
  =950+900
  =1850

D3(6750,3600) and P (7100,4000)

- |x1-x2|+|y1+y2|
  =|6750-7200|+|3600-4000|
  =1050+400
  =1450

Step 4: Now compare the D1 (4800,3750), D2 (7100,4000) values & choose the least values then form the clusters.
D1= (4450,5300) (4800,3750) (2970,6120)
D2= (7700,4500) (7100,4000) (6750,3600)

Now find the total cost in Cluster
Σ|xi| = D1=1900+4200 =6100
D2=600+1450=2050
Total Cost = 6100+2050=8150

Step 5: Now compare the D1 (4800,3750), D3 (6750,3600) values & choose the least values then form the clusters.
D1= (4450,5300) (4800,3750) (2970,6120)
D3= (7200,4500) (7100,4000) (6750,3600)
Now find the total cost in Cluster
Σ|xi| = D1=1900+4200 =6100
D3=1850+1450=3300
Total Cost = 6100+3300=9400

Now compare clusters obtained in step 4 & 5
9400>8150 so z=9400-8150=1250>0 so clusters are perfect.
Step 6: If z<0 then swap the initial points
repeat the step3,4,5,6. Until it comes to same clusters.

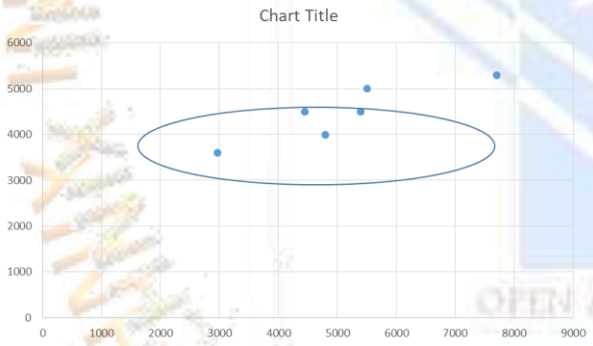


Figure: - Graph indicate points are non-associated items are not part of cluster

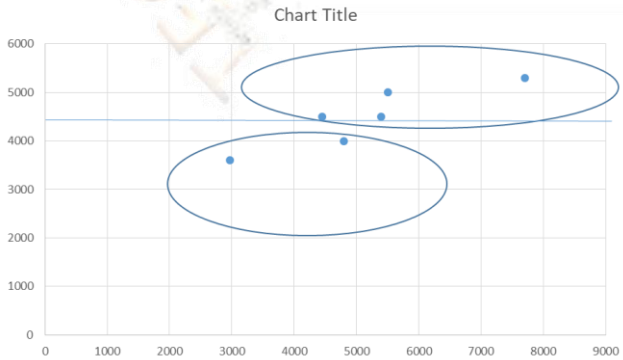


Figure: Clusters formed with non-associated cluster point added to above threshold value

In above graph originally we will not consider any of the orange points when simple clustering associated items are used, our algorithm CNAIS will show as high prioritize cluster above threshold value so that above one of the orange point can be considered as valued customer if that point belongs customer transaction data set.

VII Conclusions

In this paper a new algorithm propose Clustering Technique of Non-Association rule Items Sets (CNIAS) over Transaction item sets and sample dataset is used to give results , which considers non-associated items which are above threshold value other than Associated items we are using dataset points which when clustered is shown above threshold value. These points may be lost if simple associative clustering is used, but our technique reduces loss of data points or rows which are missed during ARM process.

References

[1] K.M.V. Madan Kumar,Dr. P.V.S. Srinivas, *Algorithms for Mining Sequential Patterns* International Journal of Information Sciences and Application. ISSN 0974-2255 Volume 3, Number 1 (2011), pp. 59-69

[2]. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207–216.

[3]Gupta, M.K., Chandra, P. A comprehensive survey of data mining. Int. j. inf. tecnol. 12, 1243–1257 (2020). https://doi.org/10.1007/s41870-020-00427-7

[4] Lent, Brian & Swami, Arun & Widom, Jennifer. (1997). Clustering association rules. Proceedings –International Conference on Data Engineering. 220-231. 10.1109/ICDE.1997.581756.

[5] S.A. Ozel, H.A. Guvenir, "An Algorithm for Mining Association Rules Using Perfect Hashing and Database Pruning" , In 10th Turkish Symposium on Artificial Intelligence and Neural Networks, Gazimagusa, T.R.N.C., A. Acan, I. Aybay, and M. Salamah, Eds. Springer, Berlin, Germany, pp.257–264, 1998.

[6] L. Padmavathy, V. Umarani, "An Efficient Association Rule Mining Using the H-BIT Array Hashing Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, vol.3, pp. 410-419, 2013.

[7] Jong Soo Park, Ming-Syan Chen and P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," in IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 5, pp. 813-825, Sept.-Oct. 1997, doi: 10.1109/69.634757.

[8] PENG Jian,WANG Xiao-ling, An Improved Association Rule Algorithm Based on Itemset Matrix and Cluster Matrix,The 7th International Conference on Computer Science & Education (ICCSE 2012),IEEE

[9] R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proc. of the 11th Int'l Conference on Data Engineering~ Taipei, Taiwan, March 1995.

[10] Ya-Han Hu a, Yen-Liang Chen, Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism, Decision Support Systems, ISSN: 0167-9236, Vol: 42, Issue: 1, Page: 1-24,2006, https://doi.org/10.1016/j.dss.2004.09.007.

[11]Man Lung Yiu and Nikos Mamoulis,Iterative Projected Clustering by Subspace Mining,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 2, FEBRUARY 2005

[12]C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, and J.S. Park, "Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD, 1999.

[13] C.C. Aggarwal and P.S. Yu, "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. ACM SIGMOD, 2000.

[14] Bellini etal.Multi Clustering Recommendation System for Fashion Retail,Multimedia Tools and Applications p1-28, 1573-7721 2022 Springer

[15] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers is an imprint of Elsevier 2006.

[16] G. Ahalya and H. M. Pandey, "Data clustering approaches survey and analysis," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, pp. 532-537, doi: 10.1109/ABLAZE.2015.7154919.

[17]Rui Xu and D. Wunsch, "Survey of clustering algorithms," in *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, May 2005, doi: 10.1109/TNN.2005.845141.

[18]Xu, D., Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci.* **2**, 165–193 (2015). https://doi.org/10.1007/s40745-015-0040-1

[19]Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7 (2002): 881-892.

[20]R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan,"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. ACM SIGMOD, 1998.

[21] M.Vinayababu, Dr. M.Sreedevi, "A Comprehensive Study on Enhanced Clustering Technique of Association Rules Over Transactional Datasets". In proceedings of IEEE fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), DOI:10.1109/I-SMAC 52330.2021. 9640681.
[22] M.Vinayababu, Dr. M.Sreedevi, "Performance Analysis on Advances in Frequent Pattern Growth Algorithm", in proceedings of IEEE International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI),DOI:10.1109/ACCAI 53970.2022.9752650.

M.Vinaya Babu, Research Scholar under the esteemed guidance of Dr.Mooramreddy Sreedevi, in the Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh. He completed MCA from Osmania University in 2001 and M.Tech(CSE) from JNTU, Hyderabad in 2013. He qualified.in APSET He attended 12 Workshops and FDPs Sponsored by AICTE and SERB DST. He published 20 research papers in UGC reputed journals and participated in 10 International Conferences. His area of interest is Data Mining. He is a Professional Member in LMISTE, MIACSIT, MIAENG, MCSTA, etc

Dr.Mooramreddy Sreedevi, working as an Associate Professor, in the Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh since 2007. She published 92 research papers in UGC reputed journals, participated in 64 International Conferences and 42 national Conferences. She acted as resource person for different universities. She acted as Deputy Warden for Women for 4 years and also acted as a EC Member and Lady Representative for 2 years in SVUniversity Teachers Association SV University Tirupati. She is acting as BOS member for UG-Computer Science of Bangalore University in Bengaluru, Karnataka, since 2018.