# Integrating Apriori with paired k-means for Cluster fixed mixed data

H. Haripriya
Amrita CREATE
Amrita Vishwa Vidyapeetham,
Amritapuri
Kollam, Kerala.India
haripriya@am.amrita.edu

Shaji Amrutha
Department of Computer Science and Application
Amrita Vishwa Vidyapeetham,
Amritapuri
Kollam,Kerala.India
mailtoamrutha01@gmail.com

R.Veena
Department of Computer Science and Application
Amrita Vishwa Vidyapeetham,
Amritapuri
Kollam,Kerala.India
veenarnair024@gmail.com

Prema Nedungadi
Amrita CREATE
Amrita Vishwa Vidyapeetham,
Amritapuri
Kollam, Kerala.India
prema@amrita.edu

## ABSTRACT
The field of data mining is concerned with finding interesting patterns from an unstructured data. A simple, popular as well as an efficient clustering technique for data analysis is k-means. But classical k-means algorithm can only be applied to numerical data where k is a user given value. But the data generated from a wide variety of domains are of mixed form and it is effortful to trust on a user given value for k. So our objective is to effectively use an association rule mining algorithm which can automatically compute the number of clusters and a pairwise distance measure for calculating the distance in mixed data. We have done experimentations with real mixed data taken from the UCI repository.

## Keywords
Clustering; Mixed data; Rule mining; Pairwise distance; Discretization.

## 1. INTRODUCTION

Knowledge extraction from structured as well as unstructured data can be effectively done with the help of various data mining techniques. This extracted knowledge helps users to analyze data in various perspectives and helps to categorize or classifies it and cluster it for further processing.

Currently there is an increasing interest of data mining in all domains, such as in Business, Bioinformatics, Education. The application of data mining in these areas will be helpful for the development of technologies and in turn the future developments. How to effectively classify or cluster a given set of data and which method will do this is the real challenge.

Classification and cluster analysis are important techniques used in data mining for grouping, that partitioned the objects into different meaningful subgroups, so that data objects within the same group are more similar to each other in the patterns of data they have than the objects outside the groups. The cluster analysis and classification has a considerable discrepancy. Classification [1] is a supervised learning in which the classes are predefined but clustering is an unsupervised technique in which it partitions the given dataset into clusters based on the identified general patterns. Simple k-means is a commonly used clustering approach due to its simplicity and effectiveness.

Classical k-means algorithm works well in numerical data sets. But in most of the real world cases we have to deal with mixed data, numerical and categorical. Another drawback of traditional k-means is the user given k value, number of clusters. Finding an appropriate number of clusters for a given data set is still remains a trial and error process and this would be difficult in deciding the exact number of clusters. The performance of the algorithm and the accuracy of the result will depend upon the selection of the k value. So our objective is to find the number of clusters in advance instead of depending on a user given value and thus reducing the manual effort.

On the other hand, the distance measure used in almost all clustering algorithm can be either used with numerical data sets or with categorical data set. We need to have a distance measure which can effectively handle both types of data. We are also modifying an already existing distance measure to a paired one for handling data from mixed domain.


*WCI '15,* August 10 - 13, 2015, Kochi, India© 2015 ACM. ISBN 978-1-4503-3361-0/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2791405.2791437

## 2. RELATED WORKS

In this section we are going through the algorithms for dealing numerical [1,3,4,5,6,7,8], categorical [9,10,11,12] and mixed data [13,14,15,16,17,18,19]. Various algorithms are helpful for dealing with each type of data. Efficient algorithms for handling mixed data is a real challenge.

### 2.1 Numerical Data

There are a variety of algorithms in data mining used for clustering numerical data. PAM [1] "Partition Around Medoids" uses a k-medoid method for clustering. Even though the goal is to minimize the average dissimilarity of objects to their closest selected object, it does not perform well with high dimensional mixed datasets.

CLARA [3] **(**Clustering Large Applications) introduced to overcome the problem of PAM. Sample data is used rather than the entire dataset and if the sample is biased the clustering will not be good.

CLARANS [4] (Clustering Large Application Based on Randomized Search) is similar to PAM and CLARA, which performs a random selection of medoids initially. It considers only a random selection of objects until it finds a better configuration rather than swaps of medoid and non-medoid objects.

One of the main drawbacks of DBSCAN [5] (Density-Based Spatial Clustering of Applications with Noise) algorithm is the need to specify global parameters Eps, MinPts in advance from user, which is very difficult. It is difficult for clustering data sets with large difference in densities.

GDBSCAN [6] (Generalized Density-Based Spatial Clustering of Applications with Noise).It is a generalized version of DBSCAN. Although it is less sensitive to outlier and form irregular shapes high dimensionality is a curse.

Fuzzy clustering[7,8] unlike other clustering algorithms the data points are mapped to one or more cluster, and each data object have a membership grade which denotes the degree of each data point belongs to different clusters. But the efficiency of the algorithm is comparatively less when dealing with real world data sets.

### 2.2 Categorical Data

Squeezer [9] is another algorithm used for clustering categorical data and can produce high quality results and have good scalability. But it does not produce accurate clusters on some data set.

Feature weighting [10] k-means algorithm is a framework for integrating more than one feature spaces in the k-means algorithm. The main disadvantage is it's inability to determine the optimal feature weighting efficiently.

In the cluster ensemble approach [11], they use a divide and conquer approach in which the data set is first classified as pure numeric and pure categorical data sets. Then the existing clustering algorithms are used for clustering the data sets. This technique would not detect the outliers effectively in the large database environment.

CACTUS [12] which is Clustering Categorical Data Using Summaries is based on fast summarization. It requires only two scans over the data set. The algorithm failed to maintain inter attribute summaries and the number of computations is larger.

### 2.3 Mixed Data

SBAC [13] (Similarity Based Agglomerative Clustering) is an unsupervised learning technique which performs well with mixed data by using a similarity measure defined by Goodall. In Goodall [14] similarity measure similarity is considered as a relation between a pair of values. Even though the algorithm is efficient for clustering mixed data it would be computationally very expensive.

Huangs [15] proposed a cost function which helps to cluster mixed type of values, it uses two algorithms k-mode and k-means. The k-prototype algorithm integrated the both algorithms for clustering the mixed valued data. It is very cost effective and handles nominal and numeric data separately. The representation of cluster centers as mode instead of mean is common in traditional means will further result in the information loss of data. It considers weight of all numeric attribute as a single value and this will result in the loss of significance of each numerical attribute.

Apriori algorithm [16, 17, 18, 19] is one of the most prominent algorithm for mining frequent item set from a large database and generating association rule for finding the number of clusters. The association rule having high confidence can be used to construct a hierarchal sequence of clusters. It can be used as a preprocessing method for finding the general pattern or most frequent pattern.

## 3. DEFINIIONS AND NOTATIONS

The notations and the definitions of terms used is explained in Table 1. Let D be the dataset D=$(o_1, o_2, \ldots o_n)$ be the set of objects in the dataset D. The mixed data set containing categorical and numerical data object is represented as,

$$D= ( o^n_{1n}, o^n_{2n}, \ldots o^n_{mn}, o^c_{m+1} \ldots\ldots o^c_{mn+1} ).$$

The notation for the distance between the data object and the cluster center is defined as, $\partial = \sum_{i=1}^{n} d(o_i, C_j)$

$$d(o_i, C_j) = \sum_{t=1}^{m_r} (s_t (o^r_{it} - C^r_{jt}))^2, \sum_{t=1}^{m_c} (\Omega(o^c_{it}, C^c_{jt})^2) \quad (1)$$

where $\sum_{t=1}^{m_r} (s_t (o^r_{it} - C^r_{jt}))^2$ represents the distance between numerical data objects $o^r_{it}$ from the cluster center $C^r_{jt}$.

**Table 1. List of symbols**

| Symbol | Definition |
|---|---|
| D | Mixed data set |
| $o_i$ | *i*-th data object |
| $o_{ij}$ | *j*-th data object of *i*-th attribute |
| $o_{mn}^{n}$ | Data object of the numerical attribute |
| $o_{mn}^{c}$ | Data object of the categorical attribute |
| $C_j$ | *j*-th cluster |
| k | Number of clusters |
| $s_t$ | Significance of *t*-th numerical attribute |
| I | Number of intervals |
| $A_i$ | *i*-th attribute of the data object |
| $\delta(x, y)$ | Distance between categorical values x and y |
| $N_c$ | Count of data object in cluster *c* |
| $N_{i,k,c}$ | Count of elements in cluster *c* which has the *k*-th attribute value for the *i*-th attribute |
| $P(x, y)$ | Probability between categorical values *x* and *y* |
| n | Count of data objects in the data set |

$\sum_{t=1}^{m_c} \Omega(o_{it}^{c}, C_{it}^{c})^2$ represents the distance between categorical data objects $o_{it}^{c}$ from the cluster center $C_{it}^{c}$.

Above described the distance measure is a modified distance measure for mixed data [20]. Min-max normalization is used as a preprocessing step in our approach to scale the data within one range.

$$o_{ij} = (o_{ij} - o_{i,\min}) / (o_{i,\max} - o_{i,\min}) \quad (2)$$

For computing the significance ( $s_t$ ) of each numerical attribute, a discretization need to be performed on attribute using equal width interval approach. Numerical data conversion is adequate for computing the significance. So first normalize the attribute and then convert it into categorical values by using equal width interval approach. Then compute the distance $\delta(v[x], v[y])$ for every pair of categorical values. The significance $s_t$ is the mean of all pairs $\delta(v[x], v[y])$, where $v[x] \neq v[y]$

$$s_t = \sum_{i=1}^{I} \sum_{j>i}^{I} \delta(v[x], v[y]) / (I(I-1)/2) \quad (3)$$

The dataset is of mixed form, i.e categorical as well as numerical data the way of representing the cluster center is different [20] from traditional approaches. The mean value is used to represent the cluster centre for numerical data and for categorical attributes the cluster center is represented as:

$$1/N_c \left\langle (N_{1,1,c}, N_{1,2,c}, \ldots, N_{i,p1,c}), .. (N_{m,1,c}, N_{m,2,c}, \ldots N_{m,pm,c}) \right\rangle \quad (4)$$

Let us consider a mixed data set with a two attributes in it's feature space, one is of type categorical attribute and another is numeric. If cluster1 contains five data objects among which three objects have *D* as value and the remaining has *E* as values for the first attribute, then the cluster center representation of cluster1 with respect to the first attribute is *{1/5(3D,2E)}*.

## 4. SYSTEM ARCHITECTURE

The overall system architecture for cluster fixed mixed data based on paired k-means paradigm is shown in *"Fig.1"*. It has mainly three phases. The first phase is for preprocessing the data and convert it into a form which could fit the algorithm. The second phase is for finding the number of clusters automatically. The last phase is for clustering mixed data based on the paired distance approach.

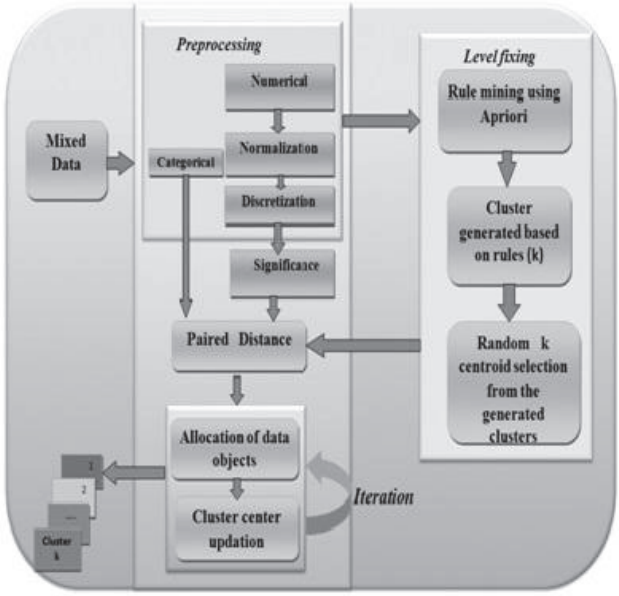


**Figure.1.System Architecture of cluster fixed paired k-means**

Input to the system is the data in mixed form and output is the clustered data. The preprocessing phase constitutes normalization

using min-max approach and discretization with equal width interval approach. All the numeric attributes of the data sets are normalized using min-max normalization before performing clustering. So the numerical data will be transformed into a range normally between 0 and 1 and which will reduce the error when dealing with continuous values. This normalized numeric attributes are then discretized using equal width interval approach for finding the significance of each of them. This significance will contribute to finding the cluster to which each object belongs.

The k value or the cluster count in k-means clustering has a significant role in the accuracy of clustering. Normally it will always depend upon the user input, but it is not a good approach to rely on a user given value. Computation of the exact number of clusters in a dataset is a challenging task. Apriori is one of the most prominent algorithms for finding frequent patterns from data. It is simple and efficient even though it is time consuming. We are using an Apriori algorithm for mining frequent item set from a large database and generating association rule for finding the cluster count. The rules with highest confidence are used for finding the possible count of clusters from a given data set.

After finding the exact number of clusters using Apriori algorithm[16] data objects from the clusters formed selected for initialization in k-means. The representation of the cluster centre is different from traditional k-means algorithm since the algorithm is dealing with categorical values also. The cluster centre for numerical attribute is the normalized mean. For numerical attribute we used euclidean distance for finding the distance between the data object and the cluster centre. The computed significance is also used in finding the distance with the numerical attribute. The centre of the cluster is represented based on each of it's attribute.

We used a pairwise distance for categorical and numerical data. A pairwise comparison is used for finding the cluster of objects. So the distance computed is a pair, one for representing the numerical values and the other for representing the categorical values rather than a single value. The remaining steps are similar to traditional k-means. An iterative process is used for finding the final set of clusters.

**Algorithm-1: Apriori_Clustering ( )**

Input - {Data: Mixed data with normalized and discretized numerical attributes}

Output - {k: Number of clusters generated}

**Begin**

1. Initialize s: = min_support; k = 0;

***Repeat***

2. R :={ Association rules over Data}
3. Best_Rule:=Rule from R having highest confidence
4. $cluster_i$ = {Objects containing products in LHS (Best_Rule)}
5. Data: =Data- $cluster_i$
6. k = k+1

***Until (Data is empty)***

**End**

The algorithm Apriori_Clustering is incorporated to generate the number of clusters for the given data set. First, it will generate a set of association rules from the data set and then, from the rules it will select the rule with higher confidence and in case if we have more rules with the same confidence, randomly select one of the rule. Based on the selected rule it scans the entire data set and choose data objects which satisfy the LHS of the selected rule and forms clusters. The process repeats until the data set becomes empty.

**Algorithm-2: Categorical_distance ( )**

Input – {Mixed data withs discretized numerical attributes}

Output-{Distance between every pair of attribute values for all attributes}

**Begin**

1. Initialize sum: =0, $\delta(x,y) := 0$

***Repeat***

{Each attribute $A_i$ in the dataset}

{Every pair of categorical attribute values $(x,y)$ of $A_i$ }

1. d (x, y):=0
2. If ( $A_i \neq A_j$ )

Repeat

If (P (u/x)>P (u/y)) then

$$\delta^{ij}(x,y) := \delta^{ij}(x,y) + P(u/x)$$

Else

$$\delta^{ij}(x,y) := \delta^{ij}(x,y) + P(u/y)$$

Until ( $A_j$ is empty)

3. $\delta^{ij}(x,y) = \delta^{ij}(x,y) - 1$
4. sum: =sum+ $\delta^{ij}(x,y)$

***Until (for all attribute other than $A_i$ )***

5. d(x,y)=sum/(m-1)

**End**

Algorithm 2, Categorical _distance () is for computing distance of categorical attributes. The probability of co-occurence of each value of a selected categorical attribute with respect to the others will be calculated. Among the computed probabilities the one with higher probability will be selected for distance computation. The process is repeated for all the values of attributes where, $A_i \neq A_j$, $A_i$ is the attribute that contains the value pair (x,y).

**Algorithm-3: Level Fixed_k-means Algorithm ( )**

Input - {D: Mixed data}

Output- {$C_1$, $C_2$...$C_k$: Clustered data}

**Begin**

1. Pre-process:
   - Separate Numerical and Categorical attributes.
   - Normalize numerical attributes.
2. Initialization:
   - k = Apriori_Clustering ( D with normalized and discretized numerical attributes )
   - Allocation of data objects to the k number of clusters randomly.

***Repeat:***

3. Compute cluster centers for each cluster.
4. Compute pairwise distance using (1)
5. Map data object to the closest cluster center based on distance.

***Until (Cluters are stable or pre-determined number of iterations are reached)***

**End**

The algorithm 3, Level Fixed_k-means is the modified k-means algorithm for dealing with mixed data. Pre-processing phase is for separating out numerical and categorical attributes and for normalization and disctreization. In the initialization phase Apriori_clustering algorithm ( ) will compute the count of clusters based on rule mining and thereby reduces manual effort. The number of clusters generated as a result of apriori is used as the k value instead of a user given value. Step 4 explains about the distance computation. The categorical_distance () algorithm is processed here for computing the distance between two categorical values.

# 5. EXPERIMENTAL EVALUATION

As in the case of classification and clustering, we can use various measures for evaluating our result. Precision, Recall, Micro-precision are all efficient measures. The accuracy and performance of the algorithms can be evaluated effectively by these measures.

Let the data set contain $C$ classes and $a_i$ be the number of data objects that are precisely mapped to the class $c_i$, $b_i$ be the number of data objects that are inaccurately mapped to the class $c_i$, and $d_i$ be the number of data objects that are incorrectly rejected from the class $c_i$ , then the equation for precision and recall of the class $c_i$ is defined as follows:

Precision of $i$-th class:

$$p_i = a_i /(a_i + b_i) \qquad 1 \le i \le C \qquad (5)$$

Recall of $i$-th class:

$$r_i = a_i /(a_i + d_i) \qquad 1 \le i \le C \qquad (6)$$

Micro-precision for the entire dataset can be calculated by averaging the over the precision value or the number of data objects that are correctly classified in the class. Similarly micro-recall is averaging over all computed recall values.

$$\text{Micro-p} = \text{micro-r} = (\sum_{i=1}^{c} a_i) / n \qquad (7)$$

## 5.1 Data Set

We have taken various real world mixed data sets from UCI repository for evaluating our result. A comparative study was done with an already existing one and our approach by making use of these datasets.

### *5.1.1 Iris*

Iris is a pure numerical data set which contains only numerical features. It has three classes, namely Iris Sets, Iris Versicolour and Iris Verginica and the data of 150 elements are equally distributed to each.

### *5.1.2 Vote*

Vote dataset contains only categorical attributes. The data set contains 435 elements distributed among 16 attributes and has two main classes Republican and Democrat.

### *5.1.3 Adult*

Adult or census income data set is a well known mixed data set in which it contains both numerical as well as categorical data. The data set contains 4882 instances among 14 attributes. The two classes are <=50K and >50K.

### *5.1.4 Credit data*

Credit data set is also a mixed data set contains numerical as well as categorical data. The data set contains a total of 690 instances among 15 attributes. The data contain data about the credit card application and has a good mix of attributes. It has mainly 2 classes + and -.

### *5.1.5 Hepatitis data*

Hepatitis data set is a mixed data set. It has both types of attributes contains categorical as well as numerical data. The data set contains a total of 150 instances distributed among 19 attributes. -Die and live are the two classes.

### *5.1.6 Horse-colic data*

It is a mixed data set contains both kinds of data. The data set contains 368 instances and 28 attributes. It has two main classes.

### *5.1.7 Teaching assistant evaluation*

This data set is a mixed data set contains both kinds of data. It is the evaluation of teaching performance as high, low and medium. It has mainly 151 instances distributed among 5 attributes.

### *5.1.8 Heart disease data set*

Heart disease data set is a mixed data set which contains the results of the evaluation of heart patients. It has 303 instances and 75 attributes and three classes.

Table 2 shows the result obtained for the Apriori algorithm, for pre-computing clusters. We have used seven real datasets for analyzing our results. The clusters generated are same for almost all datasets except two. In three cases, such as Iris, Australian credit and Teaching assistant evaluation a slight change can be seen on the number of original classes and the obtained classes, there is not a drastic difference.

**Table 2. The number of clusters obtained using association rule mining for different data sets**

| Data set | Clusters obtained | Original class |
|---|---|---|
| Vote data set | **2** | **2** |
| Credit data set | **2** | **2** |
| Hepatitis data set | **2** | **2** |
| Heart data set | **4** | **4** |
| Iris data set | 4 | 3 |
| Australian credit data set | 3 | 2 |
| Teaching assistant evaluation | 5 | 3 |

Table 3 shows the precision computed for our approach and the existing k-means for mixed data [20]. The result obtained using paired distance is comparatively higher for mixed data. So based on analyzing the obtained result, we can understand that our paired cluster fixed approach outperforms for k-means for mixed data.

**Table 3. Micro-Precision computed for Paired k-means and k-means mixed**

| Data Set | Micro-p | |
|---|---|---|
| | **Paired k-means** | **k-means mixed** |
| Iris | **0.96** | 0.88 |
| Vote | **0.88** | 0.87 |
| Adult | **0.75** | 0.72 |
| Credit | **0.71** | 0.69 |
| Horse-Colic | **0.63** | 0.60 |
| Teaching Assistant evaluation | **0.75** | 0.73 |

## 6. CONCLUSION

Classical k-means is an unsupervised clustering algorithm which is simple and efficient. Although it can resolve the clustering problems for numerical data set based on a user defined number of clusters. But considering the real world datasets most of them are in mixed form, contains numerical as well as categorical data. The accuracy of clustering depends upon the user given k value.

Sine finding the exact number of clusters for a given data set is a user driven and daunting task which in turn leads to inaccurate and bad clusters, we have integrated association rule mining algorithm for finding the number of clusters in advance and a paired distance approach in order to improve the accuracy of clustering mixed data. An elucidate analysis of the result of our approach helps us to ensure that the proposed one is better and producing accurate results compared to other approaches.

## 7. ACKNOWLEDGMENT

This work derives inspiration and direction from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. We are also expressing our gratitude to Dr. M.R Kaimal, Chairman, Department of Computer Science and Engineering and the faculties Computer Science and Application Department for their constant support and guidance. Also, we are grateful to DeitY for the support of eGAP project.

## 8. REFERENCES

1. L. Kaufman and P.J. Rousseew, 1990, *Finding Groups in Data:an Introduction to Cluster Analysis*, John Wiley & Sons, Print ISBN: 9780471878766.
2. Archana, May 2014, *Study and Comparison of Partition Based and Hierarchical Clustering*, http://www.ijarcsse.com/docs/papers/Volume_4/5_May2014/V4I5-0299.pdf, available July 20, 2015
3. Huilan Luo, Fansheng Kong and Yixiao Li, 2006, Clustering Mixed Data Based on Evidence Accumulation, Advanced Data Mining and Applications, Lecture Notes in Computer Science, ISBN 978-3-540-37026-0, pp. 348-355.
4. S.Vijayarani and S.Nithya, 2011, An Efficient Clustering Algorithm for Outlier Detection. International Journal of Computer Applications 32(7):22-27.
5. R. Ng and J. Han, 1994, Efficient and effective clustering method for spatial data mining, *Proceedings of the 20th International Conference on Very Large Data Bases,Santiago, Chile*.
6. J. Sander, M. Ester, H.-P. Kriegel and X. Xu,1998, Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, *Data Mining and Knowledge Discovery*, pp. 169–194.
7. J.C. Dunn,1974, *Some recent investigations of a new fuzzy Partitional algorithm and its application to pattern classification problems Journal of Cybernetics* 4, 1–15.
8. J.C. Bezdek, 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, *Plenum Press*, New York.

9. Z. He, X. Xu and S. Deng, 2002, Squeezer: An efficient algorithms for clustering categorical data, *Journal of Computer Science and Technology*, Volume 17, Issue 5 , pp. 611-624.

10. D.S. Modha and W.S. Spangler, 2003, *Feature weighting in k-mean clustering,* Machine Learning 52 (3) 217–237.

11. M. V. Jagannatha Reddy and B. Kavitha, 2012, Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method, *International Journalof Database Theory and Application*, Mar2012, Vol. 5 Issue 1, p121.

12. V. Ganti, J.E. Gekhre and R. Ramakrishnan, 1999, CACTUS-clustering categorical data using summaries, *in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 73–83.

13. C. Li and G. Biswas, 2002, "Unsupervised learning with mixed numeric and nominal data", IEEE Trans. Knowl. Data Eng., vol. 14, no. 4, pp. 673 -690.

14. Goodall, D., 1966. A new similarity index based on probability, Biometrics 22, pp. 882–907.

15. Huang, Z. 1997, Clustering large data sets with mixed numeric and categorical values. Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore: World Scientific, pp. 21–34.

16. Walter A. Kosters, Elena Marchaori and Ard A.J.Oerlemans,1999, *Mining clusters with association rules*, Advances in Intelligent Data Analysis, Lecture Notes in Computer Science, Springer, pp 39-50.

17. Jiao Yabing, 2013 , Research of an Improved Apriori Algorithm in Data Mining Association Rules, *International Journal of Computer and Communication Engineering* vol. 2, no. 1, pp. 25-27 , 2013.

18. Rakesh Agarwal and Ramakrishnan Srikant, 1994, *Fast algorithms for Mining association rules,* VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499.

19. Shweta and Kanwal Garg, 2013, *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, Volume 3, Issue 6, June 2013, pp. 306-312.

20. Ahmad A. and Dey L, A k-mean clustering algorithm for mixed numeric and categorical data, 2007, Data Knowl. Eng. 63, pp. 502–527.