Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Scalable Clustering Using Rank Based Pre-processing Technique for Mixed Data Sets Using Enhanced Rock Algorithm

[1]P. Parameswari*, [2]J. Abdul Samath, [3]S. Saranya

[1, 3] Department of MCA, Kumaraguru College of Technology

[2] Department of Information, Sri Ramakrishna Institute of Technology, Tamil Nadu, India

Abstract: ***The current requirements to cluster real world data sets are scalability, ability to handle any kind of data like categorical and numerical. It should also have the capability to handle noisy and missing data. Traditional algorithm can cluster categorical or numerical data but not the both. In general it is tedious to cluster mixed data types but it gives us best clusters with more accurate results. Another important factor that affects the quality of clusters are pre-processing techniques. In order to meet out the current requirement we proposed a clustering methodology that helps to enhance the performance of ROCK clustering algorithm which is scalable. This approach has two process (1) Numerical attributes are converted in to categorical, missing values are filled by using a rank based method (2) Clustering takes place using ROCK algorithm. These approaches are combined together and known as EROCK algorithm. Experimental results obtained by this methodology are compared with EM and CLOPE algorithms .It shows that our new methodology performs well for real world data sets and found it is very effective.***

Keywords: ***Clustering, Categorical data, ROCK algorithm, Mixed datasets, Data Mining***

I. INTRODUCTION

Data mining (DM) is a technology that incorporates the application of statistical techniques combined with mathematical formulae that attempt to identify considerable relationship between variables in historical data. Some of the common methodologies that make up the world of data mining include clustering, classification; association analysis etc., Data mining algorithms are rigid about data, with improved data integrity, less data redundancy and smaller correlation between attributes [1]. It is not easy to meet the requirements of data mining algorithms due to irrelevant information in actual data, which badly disturbs the competence of data mining algorithms and deviates the data mining outcome. The main goal of data mining or knowledge discovery from databases is the extraction of potentially correct information by cautious processing and analysis of data in a computationally efficient and sometimes interactive manner [2].

The main task of data pre-processing is to systematize the original business data with the new business model which clears the attributes that is inappropriate, in some situations the results became inaccurate because of inadequate data. Data is normally dirty, incomplete and inconsistent due to noisy data, errors and outliers .It also consists of discrepancies in codes or names .If the quality of data was very poor it leads to incorrect decisions and analysis. Data Warehouse (DW) also needs constant integration of quality data. Multi-Dimensional measure for data quality relies on accuracy, completeness, consistency and more. Data pre-processing includes data cleaning, filling in missing value, identify, remove outliers and resolve inconsistencies. The required data is not available for several attributes. Missing values can be due to system fault, conflicts or some other data. There are some methods to handle noisy data like binning, clustering etc, so it is essential for the user to select appropriate algorithm. Most of these algorithms fail Unsupervised learning or clustering schemes, does not make any assumptions about the category structure. Their intent is to define similarity or dissimilarity among objects. The goal of clustering is to get the structure in the form of groups where the objects are more similar. Some clustering algorithm fit for some types of applications .In few cases these algorithms can fail to cluster the data correctly when the data contains clusters of sundry shapes, densities, and sizes. Clusters can be of well-separated, center-based, contiguous, density-based, shared property or conceptual clusters**.** The clustering algorithm can be separated into five general categories [12] like hierarchical, partition clustering, grid-based clustering and others . Hierarchical clustering builds a cluster hierarchy or a tree of cluster, which is called dendrogram. Every cluster node contains child cluster and sibling cluster. Hierarchical clustering methods are grouped as agglomerative clustering and divisive clustering [12].In agglomerative clustering method each object is singleton cluster and recursively merges two or more appropriate clusters.

Normally ROCK algorithm is used to cluster only categorical dataset, cannot be used for mixed datasets. The proposed method can perform well for both numerical and categorical data sets, by means of converting numeric to categorical data. A rank based pre-processing technique is also used to handle missing values which helps to improve the cluster quality.

II. RELATED WORKS

Clustering algorithms have also been extensively studied in data mining [3]. Data clustering first appeared in the title of a 1954 article dealing with anthropological data. Data clustering is also known as Q-analysis, typology, clumping, and

taxonomy [4].K-means is a greedy algorithm which can cover from a local minimum to global optimum when clusters are well separated [5]. ROCK algorithm is used to find out the clustering for categorical data. Which is used for both real-life and synthetic data sets. For data with categorical attributes, that ROCK generates better quality clusters than traditional algorithms and exhibits good scalability properties [6]. A link based approach which effective prune outliers well in ROCK [7].

If dataset is very large, ROCK removes data size by using random sampling technique. ROCK algorithm is better than other algorithms because of implementing outlier removal scheme [8]. Pre-processing technique for handling missing value is Rank based hierarchical clustering algorithm which has comparative quality as other sophisticated algorithms and therefore are applicable to large datasets.[9] .Clustering algorithms have been used in a variety of fields such as chemistry, engineering, medicine, etc [10] and the first work related to categorical clustering is K-MODES [10], which is an extension to K-means algorithm which is represented by a centroid , and contains frequent value for each attribute and connection of the unlabeled data point calculated by the overlap distance [11].The clustering is an important task in data mining and Information retrieval[12] which is used in wide range of applications[13] .The EM (Expectation-Maximization) algorithm is an iterative clustering technique[14].It starts with an initial clustering model for the data, it refines the model to be in shape for better quality data. EM keeps the entire data into main memory and it is not scalable, the work proposed in [16] presents clustering algorithm based on k-means by introducing cost function, which works well for mixed data types. Clustering based on hierarchical clustering concepts were proposed in [for] handling data with categorical values.

III. METHODOLOGY

A. Converting Numerical Data to Categorical Data

Invariably, the real world data is either numerical or categorical in nature. While the numerical data is continuous, the categorical data comprises a set of categories. Enormous databases usually contain a mixed set of attributes which needs efficient methodology to handle their datasets. A fuzzy based clustering technique has been introduced for mixed data types, as it has been tested merely with synthetic dataset (Do ring et al, 2004). Precisely, a conversion technique to convert NTC is been proposed.

Steps for the conversion Method:

i. Scan the dataset and select the numerical attribute.

ii. Let , a_i is the numerical attribute column where, i=1,2,3….n

iii. Select the numerical attribute a_i and find the highest and least value of that attribute

iv. Find out the average of highest and least numerical attribute value Res_i

$$Res_i = \frac{Highest\ numerical\ \ value + least\ \ numerical\ \ value}{2}$$

Res_i contains an average value for every attribute column

v. Check whether every single value present in the attribute is greater than Res_i, if the condition is true consider the value as Above Res_i else take it as Below Res_i

Table 1: Dataset with numerical attribute values (Before Conversion)

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
Sunny	85	85	False	No
Overcast	65	65	True	yes
Rainy	70	70	True	No
Rainy	91	96	True	No
Overcast	81	75	False	No

Table 2: Dataset with categorical attributes values (After Conversion)

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
Sunny	Above 74.5	Above 80.5	False	No
Overcast	Below 74.5	Below 80.5	True	Yes
Rainy	Below 74.5	Below 80.5	True	No
Rainy	Below 74.5	Above 80.5	True	No
overcast	Above 74.5	Below 80.5	False	No

B. Handling Missing values

Handling Missing Values is an important phase in preprocessing for clustering operation. According to [24] 5-15% of missing data requires sophisticated methods to handle them and above 15% severely impact on decision making. Many methods are there for handling missing data, discarding and ignoring data, Maximum likelihood procedure and imputation techniques [24]. Some popular missing data handling methods are mean or mode substitution which replaces all missing values with mean and mode, but by applying this, performance of the algorithm will be affected. Regression substitutions are based on the assumption of linear relationship between the attributes. In hot deck imputation method values are replaced by a observed value of the attribute which was chosen randomly and KNN imputation replaces the missing values with K-nearest neighbor algorithm. Handling missing values is the most crucial facet of preprocessing to perform clustering operation. Therefore, a fresh rank based approach is used to handle the missing data. Let us consider a weather dataset with the following attributes.

Table 3 : Actual Weather data

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
sunny	hot	High	false	No
overcast	cool	Normal	false	Yes
rainy	cool	Normal	true	No
sunny	mild	Normal	true	Yes
sunny	cool	Normal	false	Yes
sunny	mild	Normal	false	Yes
ovarcast	mild	High	true	Yes

The processing of this rank based approach is stated subsequently.

(i) Count the number of attributes which is in a given dataset D and assign a numerical value for each attribute (1, 2…..n)

(ii) Select the first attribute$_1$ and find the occurrence of each items in that attribute (Number of times it occur) and rank them based on their occurrences. Items which occur more number of times get Rank 1 and the item which has the next higher count acquires Rank 2.

(iii) If there is any missing value in attribute1, replace it with the item which gets Rank 1.

(iv) Select the next attribute, attribute2 and rank them based on their occurrences and replace the missing values with the item which gets Rank 2 in that attribute.

(v) If two items with similar attributes get the same Rank then the item which appears primarily in that attribute will be used for replacing the missing values.

(vi) Odd numbered attribute gets Rank 1 and even numbered attributes gets Rank2.

(vii) This process continues until all the attributes are filled with missing values.

Table 4: Weather dataset with missing values

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
sunny	hot	High	false	No
overcast	cool	Normal	false	Yes
rainy	-	Normal	-	No
sunny	mild	Normal	true	Yes
sunny	cool	Normal	false	Yes
-	mild	-	false	Yes
overcast	mild	High	true	Yes

Table 5: Weather data after replacing missing values

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	High	false	no
overcast	cool	Normal	false	yes
rainy	**cool**	Normal	**true**	no
sunny	mild	Normal	true	yes
sunny	cool	Normal	false	yes
sunny	mild	**Normal**	false	yes
ovarcast	mild	High	true	yes

C. Clustering by Rock

ROCK stands for RObust Clustering using linKs [5]. It is an agglomerative hierarchy clustering which uses the links to measure similarity between data point. Initially each tuple is assigned as a separate cluster. Clusters are combined based on the closeness between clusters. Closeness is measured by finding the sum of the links between all pair of tuple. It is suitable for boolean and categorical data. In conventional approaches, categorical data are treated as boolean value. Scalability of the algorithm depends on the sample size. The Criterion function and goodness measure used is given in Eq.(1)[6] and Eq.(2)[6].

Criterion Function:

$$E_l = \sum_{1=1}^{k} ni * \sum_{p_q\, p_r \in c_i} \frac{link\,(p_q\, p_r\,)}{n_i^{1+2f(\theta)}} \qquad \text{Eq(1)[6]}$$

Where p_q,p_r represent the two points in cluster and c_i represents the i^{th} cluster and n^i represent the size of the i^{th} cluster

Goodness Measure:

$$g(c_i c_j) = \frac{link\ |c_i c_j\ |}{\left(n_{i+n_j}\right)^{1+2f(\theta)} - n_i{}^{1+2f(\theta)} - n_j{}^{1+2f(\theta)}} \qquad \text{Eq(2)[6]}$$

Methodology:

1. Draw a random sample
2. Compute the link similarity
3. Cluster with the link
4. Label it on the disk

Algorithm:

(i) Let the dataset S and K are the input for clustering. S is the dataset that comprises n number of data points which are drawn randomly from the original dataset and K is for desired number of clusters.

(ii) Consider each data point as a separate cluster.

(iii) Compute the number of links between every pair of points which is present in the dataset S.

(iv) Common neighbours between data points p_i and p_j is represented by link (p_i,p_j). By using link measure, merge the pair of data points which show more similarities.

Steps to Compute Link

a) Consider every data point as a separate cluster and set the value of each data point link as zero

b) Compute the neighbor list for all n sample points. Now each pair will have a single link.

c) The process is repeated for each and every pair of points and the link count is incremented for each pair of its neighbor's. Finally, the link counts for all pairs of points will be obtained.

(v) For each cluster i, a local heap q[i] is built and the heap is maintained during the execution of the algorithm. The cluster j in q[i] is assumed in decreasing order of the goodness measure with respect to i. Goodness measure g(i ,j) is used to find out the best two clusters.

(vi) Local heap q[i] for each cluster i maintains an additional global heap Q that contains the entire cluster. The cluster in Q is assumed in decreasing order of their goodness measure. At each step, the max cluster j in Q and the max cluster in q[i] are the best pairs of the clusters to be merged. This process is continued until k remains in cluster Q. Moreover, it also stops clustering, if the number of links between each pair of the residual cluster becomes zero.

(vii) The max-cluster u is extracted from Q and q[u] determines the best cluster v. The q[u] is a local heap and v will be merged and both u and v are deleted from the global heap Q.

(viii) The two local clusters are merged, say w and the local heap is updated.
(ix) This process is to be continued until the goal is achieved
(x) This enhancement has produced good results while the EROCK is compared with other traditional algorithms.

IV. EXPERIMENT AND RESULTS

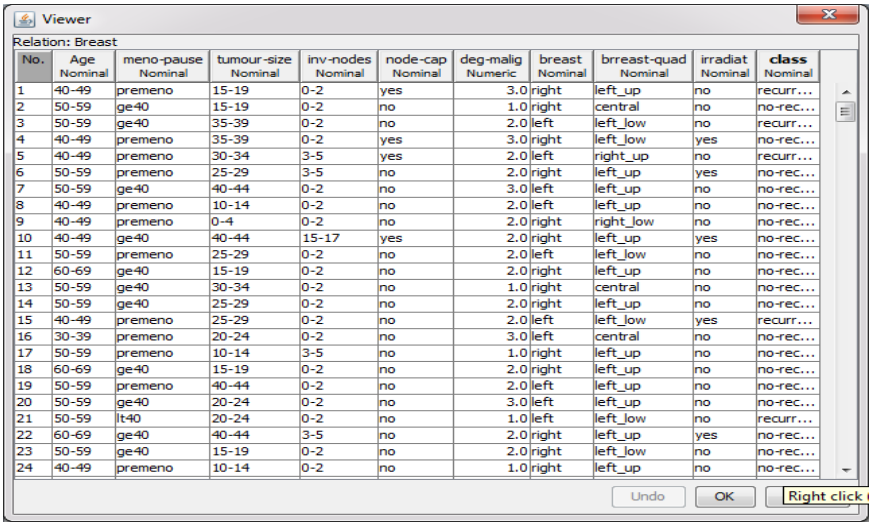
WEKA is a famous data mining tool .It is a product of the University of Waikato (New Zealand) .The software is written in JAVA language by using this tool we can perform pre-processing ,classification ,clustering etc ,it can analyse the data and produce visual results .Here we have used clustering algorithms of WEKA for comparison. The functions are inherited and used to create a tool for our research.

4.1 Performance evaluations on EROCK algorithm

In this section we evaluate Rock on categorical as well as numerical attributes The information of benchmark data sets are summarized in Table 6. We choose real world data sets from UCI Machine Learning Repository. Descriptions of data types taken for our research are given in table 1.

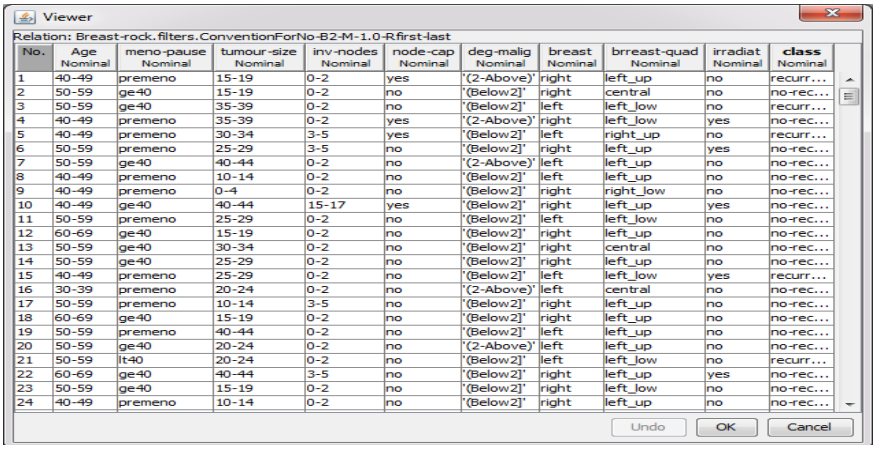
Table: 6 Breast cancer dataset

No of Record	**Data Set**	**No of Attribute**	**Missing Value**	**Note**
1500	Breast-cancer	10	Yes	453 (R), 1047 (NR)



No.	Age Nominal	meno-pause Nominal	tumour-size Nominal	inv-nodes Nominal	node-cap Nominal	deg-malig Numeric	breast Nominal	brreast-quad Nominal	irradiat Nominal	class Nominal
1	40-49	premeno	15-19	0-2	yes	3.0	right	left_up	no	recurr...
2	50-59	ge40	15-19	0-2	no	1.0	right	central	no	no-rec...
3	50-59	ge40	35-39	0-2	no	2.0	left	left_low	no	recurr...
4	40-49	premeno	35-39	0-2	yes	3.0	right	left_low	yes	no-rec...
5	40-49	premeno	30-34	3-5	yes	2.0	left	right_up	no	recurr...
6	50-59	premeno	25-29	3-5	no	2.0	right	left_up	yes	no-rec...
7	50-59	ge40	40-44	0-2	no	3.0	left	left_up	no	no-rec...
8	40-49	premeno	10-14	0-2	no	2.0	left	left_up	no	no-rec...
9	40-49	premeno	0-4	0-2	no	2.0	right	right_low	no	no-rec...
10	40-49	ge40	40-44	15-17	yes	2.0	right	left_up	yes	no-rec...
11	50-59	premeno	25-29	0-2	no	2.0	left	left_low	no	no-rec...
12	60-69	ge40	15-19	0-2	no	2.0	right	left_up	no	no-rec...
13	50-59	ge40	30-34	0-2	no	1.0	right	central	no	no-rec...
14	50-59	ge40	25-29	0-2	no	2.0	right	left_up	no	no-rec...
15	40-49	premeno	25-29	0-2	no	2.0	left	left_low	yes	recurr...
16	30-39	premeno	20-24	0-2	no	3.0	left	central	no	no-rec...
17	50-59	premeno	10-14	3-5	no	1.0	right	left_up	no	no-rec...
18	60-69	ge40	15-19	0-2	no	2.0	right	left_up	no	no-rec...
19	50-59	premeno	40-44	0-2	no	2.0	left	left_up	no	no-rec...
20	50-59	ge40	20-24	0-2	no	3.0	left	left_up	no	no-rec...
21	50-59	lt40	20-24	0-2	no	1.0	left	left_low	no	recurr...
22	60-69	ge40	40-44	3-5	no	2.0	right	left_up	yes	no-rec...
23	50-59	ge40	15-19	0-2	no	2.0	right	left_low	no	no-rec...
24	40-49	premeno	10-14	0-2	no	1.0	right	left_up	no	no-rec...

Figure 1: Breast Cancer Dataset before converting numerical values to categorical values



No.	Age Nominal	meno-pause Nominal	tumour-size Nominal	inv-nodes Nominal	node-cap Nominal	deg-malig Nominal	breast Nominal	brreast-quad Nominal	irradiat Nominal	class Nominal
1	40-49	premeno	15-19	0-2	yes	'(2-Above)'	right	left_up	no	recurr...
2	50-59	ge40	15-19	0-2	no	'(Below2]'	right	central	no	no-rec...
3	50-59	ge40	35-39	0-2	no	'(Below2]'	left	left_low	no	recurr...
4	40-49	premeno	35-39	0-2	yes	'(2-Above)'	right	left_low	yes	no-rec...
5	40-49	premeno	30-34	3-5	yes	'(Below2]'	left	right_up	no	recurr...
6	50-59	premeno	25-29	3-5	no	'(Below2]'	right	left_up	yes	no-rec...
7	50-59	ge40	40-44	0-2	no	'(2-Above)'	left	left_up	no	no-rec...
8	40-49	premeno	10-14	0-2	no	'(Below2]'	left	left_up	no	no-rec...
9	40-49	premeno	0-4	0-2	no	'(Below2]'	right	right_low	no	no-rec...
10	40-49	ge40	40-44	15-17	yes	'(Below2]'	right	left_up	yes	no-rec...
11	50-59	premeno	25-29	0-2	no	'(Below2]'	left	left_low	no	no-rec...
12	60-69	ge40	15-19	0-2	no	'(Below2]'	right	left_up	no	no-rec...
13	50-59	ge40	30-34	0-2	no	'(Below2]'	right	central	no	no-rec...
14	50-59	ge40	25-29	0-2	no	'(Below2]'	right	left_up	no	no-rec...
15	40-49	premeno	25-29	0-2	no	'(Below2]'	left	left_low	yes	recurr...
16	30-39	premeno	20-24	0-2	no	'(2-Above)'	left	central	no	no-rec...
17	50-59	premeno	10-14	3-5	no	'(Below2]'	right	left_up	no	no-rec...
18	60-69	ge40	15-19	0-2	no	'(Below2]'	right	left_up	no	no-rec...
19	50-59	premeno	40-44	0-2	no	'(Below2]'	left	left_up	no	no-rec...
20	50-59	ge40	20-24	0-2	no	'(2-Above)'	left	left_up	no	no-rec...
21	50-59	lt40	20-24	0-2	no	'(Below2]'	left	left_low	no	recurr...
22	60-69	ge40	40-44	3-5	no	'(Below2]'	right	left_up	yes	no-rec...
23	50-59	ge40	15-19	0-2	no	'(Below2]'	right	left_low	no	no-rec...
24	40-49	premeno	10-14	0-2	no	'(Below2]'	right	left_up	no	no-rec...

Figure 2 : Breast Cancer Dataset after converting numerical values to categorical values

4.2. Experiments on categorical data sets

First the investigation was done to analyse the performance of EROCK algorithm on purely categorical data. The information of benchmark data sets are summarized in Table 6. A comparison study has been made between the existing ROCK and proposed ROCK algorithm in terms of outliers d listed in Table 7 .The proposed clustering method has viable advantage in terms of accuracy and robustness, while it is compared with the existing method. Only categorical attributes are considered for evaluation and numerical attribute values for analysis are considered.

Table.7: Comparison of EROCK with ROCK in terms of Categorical and Missing values

No of Records	Rock Algorithm	Proposed ERock Algorithm
	Outlier	Outlier
150	35	4
300	37	8
450	40	15
600	45	20
750	48	26
900	50	28
1050	54	30
1200	50	35
1350	57	41
1500	63	45

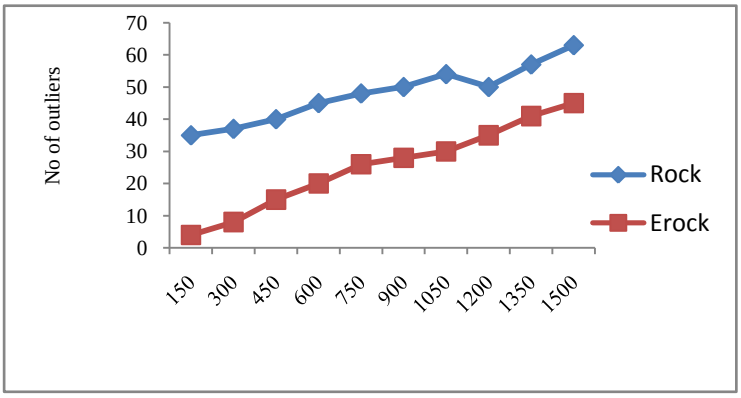


Figure 3 : Comparison between ROCK and EROCK

4.3. Performance evaluation of EROCK algorithm on mixed datasets

To examine the efficiency of the proposed method, real data sets, including purely categorical data and mixed data are considered for evaluation. EM algorithm and CLOPE algorithm are taken for comparative study showing that EROCK out performs all other clustering algorithms and performed well on mixed datasets. The results of the selected datasets are shown in table 8. This dataset has a mixed data type (categorical and numerical),it also have some missing values.

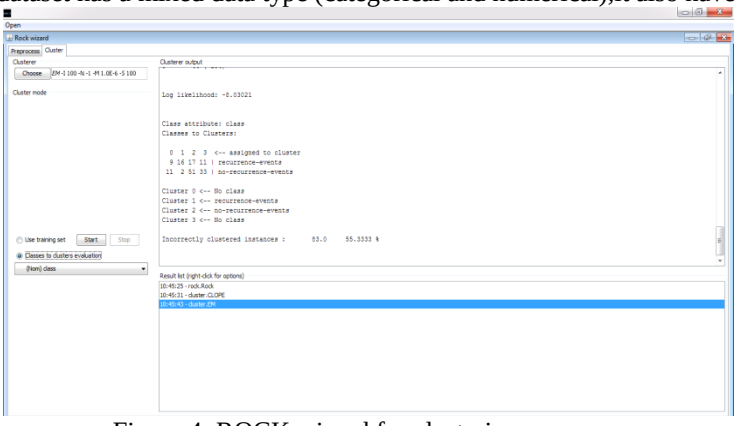


Figure 4: ROCK wizard for clustering

Table 8: Comparison of the clustering errors obtained by three different methods on mixed data sets.

Instances	Rock Algorithm		EM Algorithm		Clobe Algorithm	
	Incorrectly cluster	Error Rate	Incorrectly cluster	Error Rate	Incorrectly cluster	Error Rate
1500	881	58.85	1158	77.2	1361	90.91

1200	708	59	990	82.5	1075	89.58
900	450	49.99	704	78.2	823	91.44
600	285	47.5	457	77.45	518	87.79
300	134	44.66	171	60	252	88.42
150	62	41.33	83	55	131	87.33

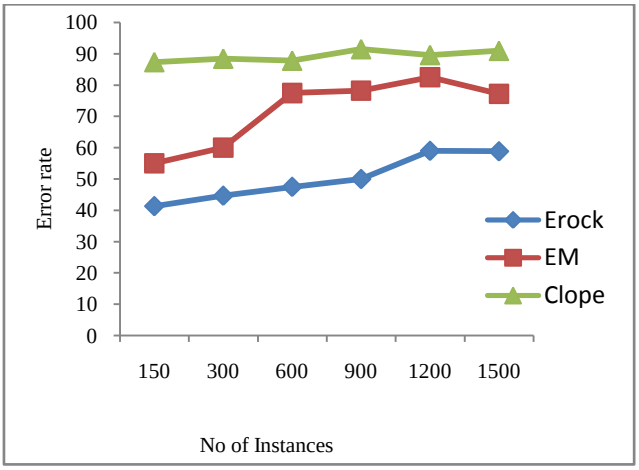


Figure 5: EROCK Vs EM & CLOPE in terms of Error Rate

Table: 9 Comparison of the execution time between three different methods on mixed data sets.

Instances	**ERock**	**Clope**	**EM**
1500	167.39	200.4	662.48
1200	106.02	150.5	509.04
900	41.7	56.6	49.82
600	5.1	12.7	41.81
300	0.34	3.5	12.25
150	0.05	2.9	3.89

The above table shows that EROCK is better than CLOPE and EM method. CLOPE performs better than EM method but EROCK outperforms other two methods.

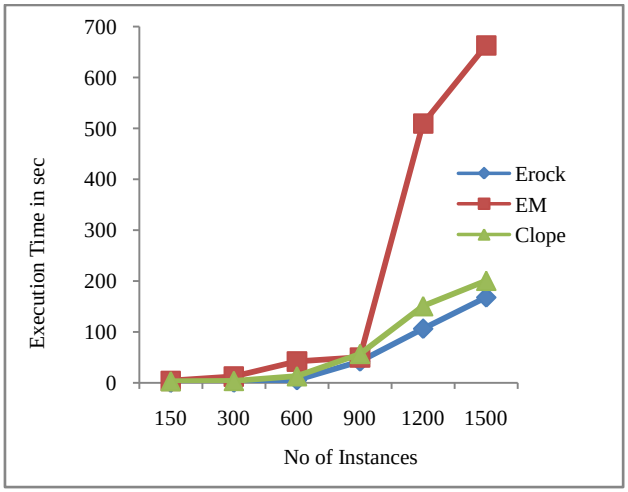


Figure 6: EROCK Vs EM & CLOPE in terms of Execution Time (sec)

V. CONCLUSION

The biggest advantage of ROCK algorithm is scalability and efficiency, but it is limited to categorical values. Therefore we proposed a approach to convert the numerical values to categorical which is a weakness of ROCK algorithm ,then we replaced the missing values by using rank based approach .The proposed approach integrated the conversion, pre-processing and clustering which helps us to provide valuable clusters. An experimental result shows that our new approach can produce high quality clustering results. This approach is applied for real world datasets and is found very effective.

REFERENCES

[1] Balaji Padmanabhan,Data Mining for Customer Segmentation: A Behavioral Pattern-Based Approach . The Wharton School, University of Pennsylvania.2004.

[2] Li.C and G. Biswas, Knowledge-based Scientific Discovery from Geological Databases,º Proc. First International Conference on Knowledge Discovery and Data Mining, pp. 204-209,1995

[3] Jiawei Han, Micheline Kamber, Jian Pei,. Data Mining: Concepts and Techniques. Morgan Kaufmann Series,2001

[4] www.jstor.org.

[5] Meila, Marina, The uniqueness of a good optimum for k-means. In: Proc. 23rd International. Conf. Machine Learning, pp. 625–632,2006

[6] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim , A Robust Clustering Algorithm For Categorical Attributes, Stanford University, Standford, CA 94305,USA,2000

[7] Sueli A. Mingoti, Renata A. Matos, .Clustering Algorithm for categorical Data. International Journal of Statistics and Applications ,pp: 24-32,2012

[8] Amir Ahamad, Lipika Dey .A k-mean clustering algorithm for mixed numeric and categorical data, Science Direct, Data & Knowledge Engineering pp: 503–527,2007

[9] Yoshikazu Fujikava and Tubao Ha "Scalable Algorithm For Dealing with Missing value" Japan Advanced Institute of Science and Technology.

[10] Huang.Z , Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min. Knowl. Discov.2(3),pp: 283-304,1998.

[11] Stanfill.C and D. Waltz, Toward memory-based reasoning. Commun. ACM,29(12),pp:1213- 1228,1986

[12] Charikar.M, C. Chekuri, T. Feder and R. Motwani, , Incremental Clustering and Dynamic Information Retrieval. SIAM J. Computing 3(6),pp:1417-1440,2004

[13] Jain A.K, M. N. Murty and P. J.Flynn, Data Clustering: A Review. ACM Comput. Surv. Vol 31(3),pp:264-323,1999

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, .Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society,1977

[15] P. Cheeseman and J. Stutz. Matthew Self, Jim Kelly ,Bayesian classification (auto class): Theory and results. AAAI -88 proceeding.

[16] Amir Ahmad , Lipika Dey , A k-mean clustering algorithm for mixed numeric and categorical data , Data & Knowledge Engineering 63 (2007) 503–527

[17] Fisher D.H, Knowledge acquisition via incremental conceptual clustering, Machine Learning 2 (2) pp:139–172,1987

[18] Li.C, G. Biswas, Unsupervised learning with mixed numeric and nominal data, IEEE Transactions on Knowledge and Data Engineering14 (4) pp:673–690,2002

[19] Goodall D.W, A new similarity index based on probability, Biometric 22 pp:882–907,1996

[20] /www.cs.waikato.ac.nz/ml/weka.

[21] He.Z, X. Xu, S. Deng, Squeezer: An efficient algorithms for clustering categorical data, Journal of Computer Science and Technology 17 (5) pp: 611–624,2001

[22] Anil K. Jain Data clustering: 50 years beyond K-means Pattern Recognition Letters 31(2010) 651–666.

[23] Yiu-ming Cheung a, b, n, Hong Jia , Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number Pattern Recognition 46 pp:2228–2238,2013.

[24] Acuna, E. & Rodriguez, C.The Treatment of Missing Values and its Effect on Classifier Accuracy, In: Classification, Clustering, and Data Mining Applications, pp: 639-647,2004

[25] UCI Machine learning Repository-Datasets.

