Validating the Effect of Different Discretization Methods for Redic K-Prototype Clustering Algorithm

Khyati R Nirmal, K.V.V. Satyanarayana

Abstract*: The REDIC K-prototype clustering algorithm is designed for mixed datasets which selects the initial centroids significantly and it also removes the dependency on prior value for number of cluster (k) and influence parameter (λ). Data preprocessing on data set introduce empirical better performance for any data mining algorithm. In this paper the taxonomy is build by integrating the data preprocessing technique – discretization with REDIC K-Prototype clustering algorithm. This taxonomy validates the performance of the algorithm for four different dataset and three performance indices. The numerical attributes of dataset need to be discretized and converted to categorical attribute before the clustering. Here the four discretization techniques are considered Equal Width Binning, Equal Frequency Binning, Entropy Based Binning, and the special case of Equal Width Binning that is binary Binning Approach. The result of proposed algorithm are compared with the standard K-Mode and K-Prototype clustering for original dataset and discretized data set. From the performance analysis it is clear that for 70% cases the REDIC K-Prototype Clustering with different discretization method gives better performance in compare to standard algorithms.

Index Terms*: REDIC K-Prototype Clustering Algorithm, Discretization; Equal Frequency Binning; Equal width Binning; Entropy based Binning

I. INTRODUCTION

today's digital word the rate of data production has been increased rapidly. It is the most common task is to analyze the data. A range of Data Mining and Machine learning algorithm lend a hand to analyze the data in effective and efficient way. To identify the group of similar data by considering the hidden structure of data , the K Means clustering and its variations are widely adopted.[1][2]

K means clustering mainly designed for numerical datasets, but in real world the data can be in any form numerical, categorical or mixed. Z.Huang has proposed the extension of K Means Clustering algorithm for categorical dataset and mixed data set. For categorical data set the K-Mode Clustering Algorithm and for mix dataset K-Prototype Clustering Algorithm has been proposed. These two algorithm works in similar manner of K Means Clustering Algorithm. The only difference is in dissimilarity measurement methods. Along with Euclidean Distance measurement formula for numerical attributes, the Simple matching dissimilarity method is used for categorical datasets. [3] However the clustering result of these algorithms are depend on major two parameters, The k value for number of clusters and initial centroid selection. Different initial cluster will result in different allotment of clusters. Some methods are proposed to improve the efficiency and stability of algorithm [4] –[8].

These methods are either computationally expensive or handle the issue of initial centroid selection or decide the number of cluster without user interference but not all of three.

Removal Dependency on K and Initial Centroid Selection (REDIC) K-Prototype clustering algorithm for mixed data has been proposed. The algorithm is computationally not much expensive and the evaluation results are also better than the K-Prototype algorithm. [9]

The result of any data mining algorithm can be improved by introducing the data processing and data transformation. One of the widely used data transformation technique is Data Discretization.

Discretization is a process of quantizing continuous attributes. The different methods are chosen for discretization based on supervised or unsupervised learning algorithm. [10]

The objective of this paper is to validate the performance of REDIC K-Prototype algorithm by integrating different binning approach of data Discretization. The paper is organized as: Section 2 has overview of binning approach of Discretization. The integration of binning approach with REDIC K-Prototype Clustering algorithm is proposed in chapter 3. In Section 4 the evaluation result of different binning approach for REDIC K-Prototype has been carried out and compare with K-Mode clustering and K-Prototype clustering algorithm.

II. OVERVIEW OF DISCRETIZATION

Discretization is to convert a continuous data attributes into categorical data attributes by considering a limited number of intervals. Discretization is integrated with many machine learning and data mining for data transformation purpose. Examples of such algorithms where discretization is adopted and have significant results are decision trees, random forests, Bayesian networks, Naive Bayes ,rule-learners, K Means Clustering [11][13]. Discretization methods are categorized in many ways.

Depending on the prior information of data set is used or not, the Discretization methods are of two types supervised or unsupervised. While doing the partition of continuous attributes the supervised methods make use of the class label. In unsupervised Discretization methods the class label information are not used. Equal Width

Revised Manuscript Received on June 05, 2019

Khyati R. NIrmal, ResearchScholar,Department of CSE, KoneruLakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh .

K.V.V.Satyanarayana, Department of CSE, KoneruLakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh.

Binning and Equal Frequency Binning are examples of unsupervised method and Entropy based binning is example of supervised method.

Discretization can be applied prior of training, or during the training. Depending on this discretization is categorized into Global and Local Method. Again when Discretization method discretize continuous values when a classifier is being built is known as Dynamic method and in static method the discretization is done prior to classification.

Another way to categorize the discretization methods is direct vs. incremental. Direct methods (Equal Width Binning and Equal Frequency Binning)divide the range of k intervals simultaneously depending on the input given by user. Incremental methods initiate with a straightforward discretization and continue throughout an improvement process, needing an additional criterion to know when to stop discretization.

This paper mainly focus on four types of discretization methods, Equal Width Binning, Equal Frequency Binning, Entropy Based Binning, and the special case of Equal Width Binning that is binary Binning Approach.

A. Equal Width Internal Discretization

The EW discretization method transforms continuous attributes to ordinal attributes. The range of values divided into k equal bins, where k is specified by user. The minimum and maximum attribute from respective attribute is selected, using these values he bin width is calculated using equation 1.

$$\text{Bin width} = \frac{a_{max} - a_{min}}{k} \qquad (1)$$

The boundary is calculated by considering Bin Width and K value using equation 2.

$$\text{Boundries} = a_{min} + (i * \text{BinWidth}) \qquad (2)$$
$$\text{where} i = 1,2,3,\ldots,k-1$$

B. Equal Frequency Interval Discretization

In EF discretization method The minimum and maximum attribute from respective attribute is selected and all the data instances for respective attribute are arranged in ascending order. Here again the data is divided into k bins, and each bin contains n/k data instances. In case of the residual instances, the last data instance is added to last bin.

C. Entropy Based Discretization

This is the supervised type of discretization method, here the entropy and information gain of the data are calculates as per the class label. The split with maximum information gain is considered as best split. Here best split is considered such that the bins are more and more promising that is the majority of the data instances in a particular bin correspond to have the same class label.

III. PROPOSED ALGORITHM WITH DISCRETIZATION

REDIC K-Prototype algorithm is proposed for mixed data which removes the dependency of prior selection of K (number of clusters) and initial centroids are also calculated automatically. The algorithm calculates the significant attributes which represent the clusters as initial centroids, and allocate the instances to the cluster having minimum distances with centroids. In proposed REDIC K-Prototype clustering algorithm for the categorical attributes the frequency based dissimilarity measurement method is used instead of bit by bit comparison method of K-Prototype clustering algorithm. The proposed REDIC K-Prototype algorithm is evaluated for raw data only, the data preprocessing in not introduced in the paper .Here the REDIC K-Prototype algorithm is integrated with discretization method of data preprocessing. This algorithm utilize the almost approach proposed in [9].

The Taxonomy for validating the performance of REDIC K-Prototype clustering algorithm is proposed in Figure 1 Firstly the original data set is discretized using different method, and the continuous attributes are converted to categorical attributes. Here the numerical attributes are considered as continuous attribute and converted to categorical attributes.

These discretized datasets are considered as input to the proposed and standard algorithm. The processed data set is having only categorical attributes, so K-Mode clustering algorithm is selected as standard algorithm.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

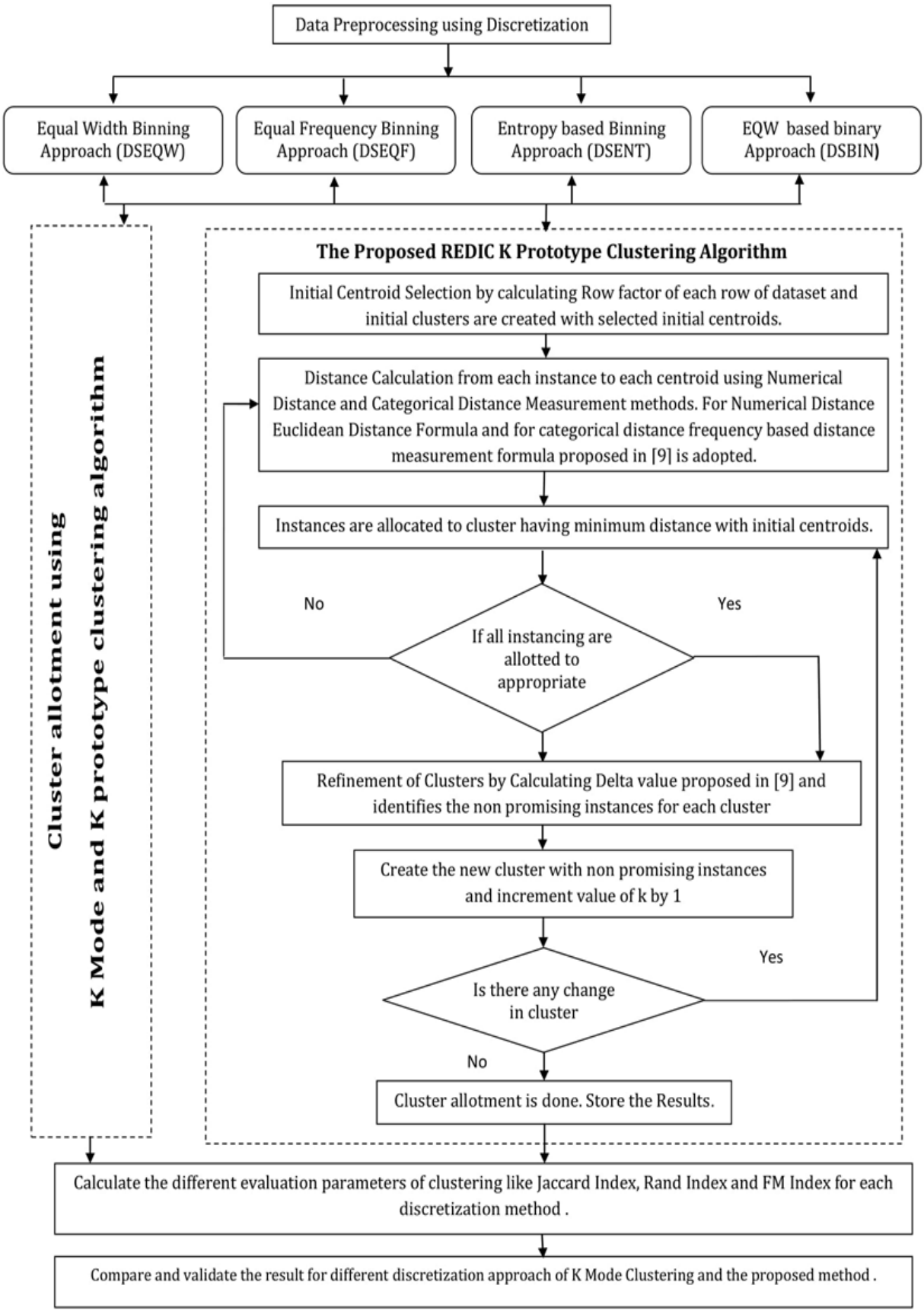

Figure 1 Taxonomy for validating the performance of REDIC K-Prototype clustering algorithm

The K-Mode clustering algorithm selects the centroids randomly, where the REDIC K-Prototype calculates the significance attributes, which will act as initial

centroids. The number of clusters is priory decided and given as input to K-Mode clustering, while in REDIC K-Prototype clustering,

The number of clusters is decided during instances allocation by means of incremental approach.

All the instances of the dataset are allotted to appropriate clusters and different performance measurement indices are calculated.

IV. RESULTS AND ANALYSIS

The section is separated into three parts, the data set used, the performance measurement indices considered and the final results.

For experimental analysis four datasets are considered: post-operative patient dataset, Australian credit dataset, German credit dataset and Statlog (Heart) dataset are used. This dataset are openly available on UCI repository. The dataset is of mixed type and basic numerical and categorical attributes distribution of particular dataset is specified in below table. Here the numerical attributes are converted into categorical attributes using different discretization techniques.

In this paper mainly three performance measurement indices are considered: Rand Index, Jaccard Index, Rand Coefficient, Folkes and Mallow index. These external indices define an evaluation measure on the basis of agreement and disagreement between object pairs in clustering. Without considering ground truth that is class label of data, it compares the results of clustering. As an outcome these indices are independent of the cluster description and can be applied for any clustering algorithm. The value of these performance indices ranges from 0 to 1. The highest value for these indices is 1,so maximum value while comparison indicates the better result.

To validate the performance of proposed algorithm REDIC K-Prototype clustering algorithm for different discretization methods, the abbreviation for different discretization methods are considered.

WDS – Without Discretization, DSEQW- Equal Width binning approach of Discretization, DSEQF - Equal frequency binning approach of Discretization, DSENT- Entropy based binning approach of Discretization DSBIN- Equal width binary approach of Discretization. The original data set and these four discretized data set act as input for proposed algorithm and performance measurements are calculated.

From Figure 2 it is predicted that the proposed algorithm provides the equivalent or better result without discretization of data for Rand Index.

In Figure 3 for Jaccard Index, DSBIN approach is providing equivalent or better result for statlog, pod and credit data set. But for crds data set the DSENT and DSEQW is providing better result.

TABLE 1. Description of Standard Dataset

Dataset	Acronym used	No of Instances	Total Attributes	Numerical Attributes	Categorical Attributes
Post-operative patient	pod	90	8	1	7
Australian credit Data Set	crds	690	15	6	9
German credit Data Set	creditg	1000	20	7	13
Statlog (Heart) Data Set	Statlog heart	270	13	9	4

In comparison of FM Index, the DSENT approach demonstrates the better result in all three dataset, except crds dataset. DSBIN approach is is also giving equivalent result for credit and statlog heart dataset.

Table 2 records the performance of proposed REDIC K-Prototype Clustering algorithm with different Discretization methods. The results are compared with the standard K-Mode and K-Prototype algorithm.

Each discretization methods acquire different characteristics and all the methods performance differently for different type of attributes.. Every discretization methods have its-own strengths. The effect of different discretized dataset for different dataset gives the superior, low-grade or equivalent results.

In 75 % cases the proposed methods are giving superior performance for all cases. But for Rand Index of crds dataset, Jaccard Index and FM Index of Statelog heart the results are equivalents or not the superior.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

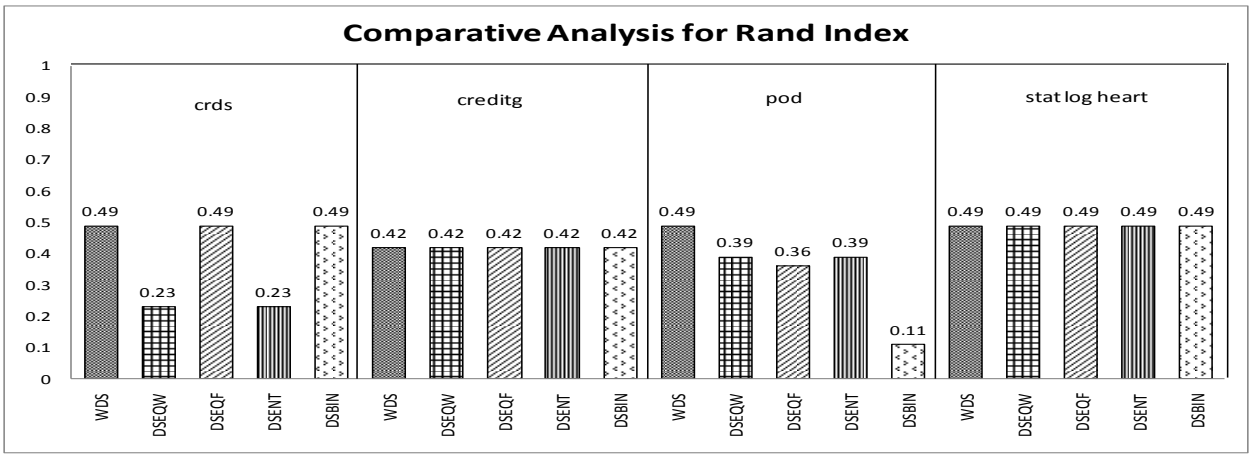

Figure 2: Comparative Analysis for Rand Index

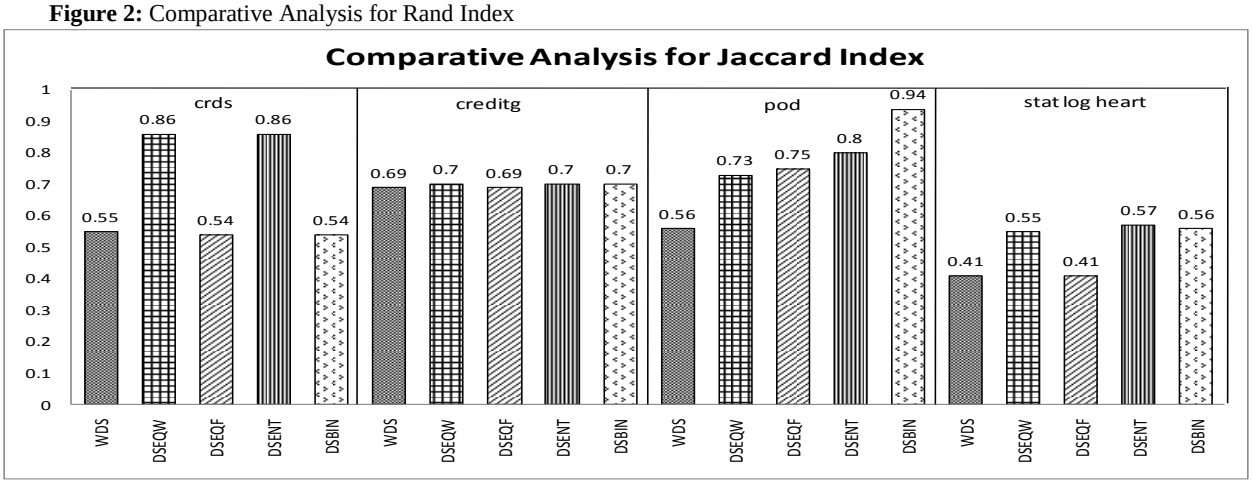

Figure 3: Comparative Analysis for Jaccard Index

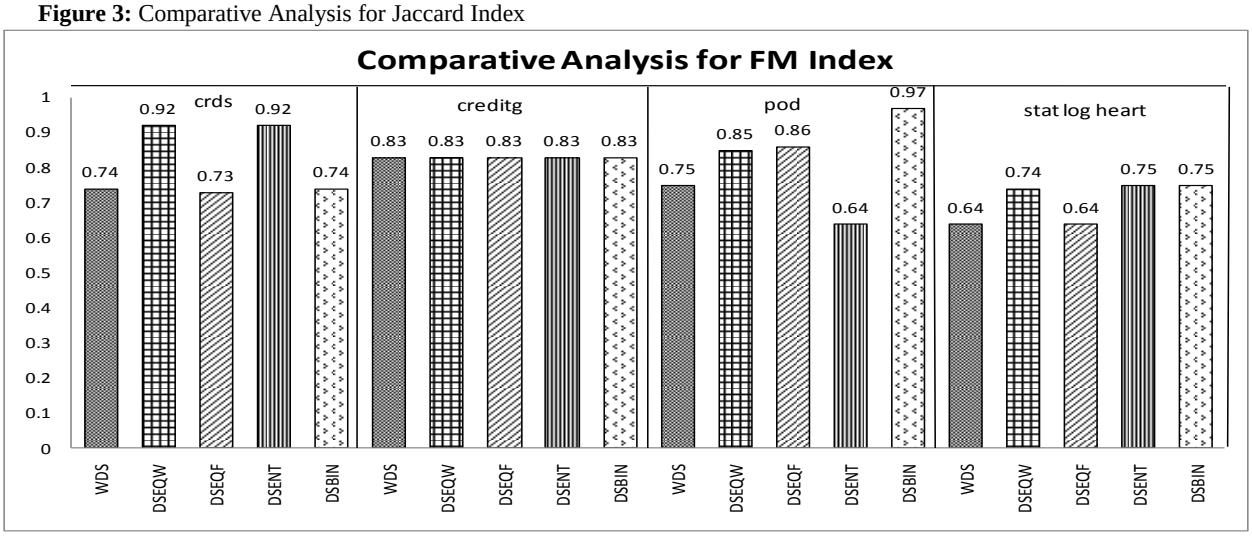

Figure 4: Comparative Analysis for FM Index

Table 2 Performance of proposed REDIC K-Prototype Clustering algorithm with Discretization methods

Dataset	Algorithm		K-Mode and K-Prototype Evauation					Proposed REDIC K-Prototype Evaluation				
	Discretization methods		WDS	DSEQW	DSEQF	DSENT	DSBIN	WDS	DSEQW	DSEQF	DSENT	DSBIN
crds	Rand Index	=	**0.49**	0.65	0.41	0.47	0.32	**0.49**	0.23	**0.49**	0.23	**0.49**

	Jaccard Index	↑	0.54	0.42	0.29	0.39	0.79	0.55	**0.86**	0.54	**0.86**	0.54
	FM Index	↑	0.73	0.49	0.54	0.62	0.89	0.74	**0.92**	0.73	**0.92**	0.74
creditg	**Rand Index**	↓	0.48	0.49	0.49	0.5	0.46	0.42	0.42	0.42	0.42	0.42
	Jaccard Index	↑	0.59	0.56	0.43	0.51	0.35	0.69	**0.7**	0.69	**0.7**	**0.7**
	FM Index	↑	0.77	0.74	0.66	0.71	0.59	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
pod	**Rand Index**	↑	0.34	0.28	0.25	0.23	0.37	**0.49**	0.39	0.36	0.39	0.11
	Jaccard Index	↑	0.43	0.71	0.64	0.42	0.55	0.56	0.73	0.75	**0.8**	0.94
	FM Index	↑	0.65	0.5	0.8	0.65	0.74	0.75	0.85	0.86	0.64	**0.97**
StatLog heart	**Rand Index**	↑	0.2	0.34	0.36	0.38	0.33	**0.49**	**0.49**	**0.49**	**0.49**	**0.49**
	Jaccard Index	↓	0.35	**0.77**	0.75	0.74	0.78	0.41	0.55	0.41	0.57	0.56
	FM Index	↓	0.59	**0.88**	0.87	0.86	**0.88**	0.64	0.74	0.64	0.75	0.75

V. CONCLUSION

In this paper the performance of the proposed REDIC K-Prototype clustering is validated for different data discretization methods. The purpose of discretization is to transfer the continuous attributes for into the categorical attributes for better performance of clustering algorithm. Here the taxonomy is built by conducting the experiments on four real data set and four different discretization methods. From the above experiments it has been found that the performance of clustering result depends on the characteristic of dataset chosen and the discretization approach considered. The main conclusion of this paper is to predict which discretization method is better for which dataset. This paper also concludes that discretization is significant stage of data mining and the selection of appropriate method will definitely helps to improve the performance of results. It will also offer the better data preprocessing practice for different machine learning and data mining algorithm. So it is significantly essential to pick appropriate discretization methods depending on data sets and learning context in practice.

REFERENCES

1. Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference. Singapore: World Scientific, 1997, pp. 21–34
2. Ahmad and L. Dey, “A k-mean clustering algorithm for mixed numeric and categorical data,” Data and Knowledge Engineering, vol. 63, no. 2, pp.503–527, 2007.
3. Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data mining and knowledge discovery, vol. 2, no. 3, pp. 283-304, 1998
4. Ji, Jinchao, et al. "An improved k-prototypes clustering algorithm for mixed numeric and categorical data." *Neurocomputing* 120 (2013): 590-596.
5. Najjar, Ahmed, Christian Gagné, and Daniel Reinharz. "A novel mixed values k-prototypes algorithm with application to health care databases mining." *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*. IEEE, 2014.
6. Xuan, Chen. "An improved clustering algorithm for mixed attributes data based on k-prototypes algorithm." *2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*. IEEE, 2015.
7. Wangchamhan, T., Chiewchanwattana, S., &Sunat, K. (2017). Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering. Expert Systems with Applications, 90, 146-167.
8. angam, R. S., & Om, H. (2018). An equi-biased k-prototypes algorithm for clustering mixed-type data. Sādhanā, 43(3), 37.
9. Nirmal K.R. &K.V.V.Satyanarayana, REDIC K prototype Clustering Algorithm for Mixed Data (Numerical and Categorical Data)International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6,pp.1-6 March 2019
10. Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. Data mining and knowledge discovery, 6(4), 393-423.
11. HACIBEYOĞLU, M., & IBRAHIM, M. H. (2016). Comparison of the effect of unsupervised and supervised discretization methods on classification process. International Journal of Intelligent Systems and Applications in Engineering, 105-108
12. Han, J., Pei, J., &Kamber, M. (2011). Data mining: concepts and techniques. Elsevier
13. Foorthuis, R.: SECODA: Segmentation- and Combination-Based Detection of Anomalies. Proceedings of the 4th IEEE International Conference on Data Science and Advanced An-alytics (DSAA 2017), Tokyo, Japan, pp. 755-764 (2017)
14. Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. International Journal of Advances in Science and Technology, 2(3), 29-37.