International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 13 (2019) pp. 2995-3002


# Cluster Sampling to Improve Classifier Accuracy for Categorical Data

**Lakshmi Sreenivasa Reddy D**
*Associate Professor*
*Department of Information Technology, Chaitanya Bharathi Institute of Technology*
*Gandipet, Hyderabad-500075, Telangana, India.*

**Abstract**

Clustering is one of the essential techniques to group similar data. Improving model accuracy is still a challenge for all variety of data. Training and testing a classifier on entire data is not possible for large scale of data. Sampling of the data is necessary for any modeling and is an important aspect in data mining. All models train and test on different samples taken by traditional techniques like random forest ensemble method. In this paper, we propose cluster sampling which is superior to any other sampling methods in improving classifier accuracy. Sampling the data from usual methods cannot cover all variety of data from the original. Cluster sampling is a two-step approach. First it clusters the entire data, second it selects samples from each cluster. These samples consists all verity of data with equal proportion. Cluster sampling leverages the tree based ensemble to handle categorical, numerical and mixed type of data. Classifiers modeled on cluster sampling samples shown superior in accuracy than modeled on other sampling techniques.

**Keywords:** Clustering, Categorical data, Numerical data, Random forest, Classifier, Sampling

## I. INTRODUCTION

Clustering algorithms group the entire data into different portions with similarity and an unsupervised learning. After grouping the data, useful patterns can be identified and characterized the groups. There are many challenges remain in clustering.

Classification is another challenge in data mining in modeling the data and getting good precision. While modeling the data we generally use traditional sampling methods. Real life data contains high dimensionality is another big challenge for clustering, specifically which are dealing with Euclidean distance measurement. When objects are equidistant in clustering the curse of dimensionality problem arises [1] and NN problem can be ill defined [2]. Many real life data consists noisy, irrelevant and redundant data which leads wrong patterns from the clusters. Distribution of the data is very important to train correct classifier. Distribution is different for different types of data. Distribution of categorical data remains an important task. Majority of data related applications deal with the categorical or mixed data.

Here we propose a sampling algorithm which improves the classifier accuracy and addresses many challenges in different aspects. Our algorithm stars with clustering and then distributing the data from clusters.

## II. RELATED METHODS

There are four types of clustering algorithms: partitioning-based, density-based, hierarchical-based and grid-based [6]. All these algorithms have been tried to handle mixed attribute and high dimensional data sets. In partition type, K-Means is one of the popular partitioning algorithms [7] that use the mean instance as the center of the cluster. K-Medoids is another well-known portioning algorithm [8] which is more robust in dealing with noise and anomalies in the data. It uses the representative instances for each cluster for next iteration instead of mean instance by selecting random instances from clusters. K- Modes is another algorithm extension to k-Means to deal with mixed and categorical data. K- Modes [9] uses frequency of attribute values and creates most frequent instance using most frequent attribute values.

This uses the simple matching dissimilarity distance measure to handle categorical data. K-modes replaces the means with modes for each cluster and the modes are updated iteration by iteration with frequency based methods. K- Prototype algorithm [9] is used to handle mixed type of data. It uses the linear combination of two dissimilarity measures to integrate K-Means for numeric and K-Modes for categorical parts of the data. More specifically the numeric attributes are measured by Euclidean distance and categorical attributes measured by simple matching. A set of suitable weights are needed to maintain balance between these two parts of data.

Agglomerative and divisive methods are two parts under hierarchical clustering. Clusters at lower level merge to large cluster iteration by iteration base on proximity measure. Number of cluster are automatically formed based on proximity measure, there is no need of giving predetermined cluster number. Complexity of Agglomerative algorithm is $O(N^3)$. This algorithm is not useful for large datasets and useful for numerical data. For categorical data, the extension of Agglomerative ROCK[10] is used. It defines the neighborhood based on the number of links between two records. If the number of records are more than or equal to a certain threshold. If the similarity exceeds the threshold they become into a cluster. ROCK does not scale well for large data. CLIQUE[12] is another algorithm to handle

dimensionality. It finds density based clusters using subspace. The clusters generated by CLIQUE are highly affected with parameter called density threshold.

PAM [8] is another best realization of K-Medoids algorithm. K-Medoid selects k medoids arbitrarily from the data set and it swaps non medoid with medoid to get optimal medoids finally. The optimal medoids are estimated using the estimated distance of all clusters between the medoids to non medoids. The slight extension of CLIQUE is the selection of medoids are arbitrary. It minimizes the expected distance as below

$$D = \sum_{i=1}^{k} \sum_{m \in Ci} sim(n, m_i) \quad (1)$$

Where n represents non-medoid and m represents medoid of the $i^{th}$ cluster. Complexity of the PAM is high. To reduce the complexity, CLARA algorithm is used based on sampling concept. CLARA uses K- Medoid concepts more applicsble for large data, but compromises the quality of clusters.

## III. CLUSTER SAMPLING

The proposed Algorithm selects different level of samples from each cluster and creates a new data set for training and testing the classifier. This sampling is like stratified sampling. Here the strata are clusters. The main logic in this algorithm is divided into two parts. First, dividing the data into clusters with high intra similarity and low inter similarity. Second, it selects the instances from each cluster with some percentage of instances so that all variety of data will be selected and grouped into big cluster. In traditional sampling, the data is selected as samples randomly. From traditional sampling there is no guarantee of selecting all variety of data.

After drawn the data from each cluster union all instances and trained and tested the classifiers CRT, CAHID,and Neural Networks.

## IV. INITIALIZATION

Dataset *D* with size *N*, first random sample of size *n* is drawn, and then divided into k- clusters using different cluster techniques. Samples of size $SM_i$ 20%, 40%, 60% are drawn from each cluster $C_i$ . *d* is the new data set formed from all $SM_i$.

Algorithm 1. Sample clustering

*Initialization (D,K,SM)*

*Call clustering algorithm for K clusters Ci*

*For* $i \leftarrow 1,K$ ***do***

*Draw the sample data* $SM_i \in Ci$

$di \leftarrow U\{SM_i\}$

*Build classifier on di*

*end*

## V. EXPERIMENTAL RESULTS

Real datasets both numeric and categorical data sets have been used for experiments in this work.

Breast Cancer data which is numeric with size 569 instances, 30 attributes and two classes Benign and Malingnant. Mice Protein data with 1080 instances, 77 attributes and two classes CS, SC. Both real data chosen for experiments are numeric data. All these data sets have been taken from UCI ML data repository. The results of Clustering, their purity and classifier's accuracies are given in the below tables with Figs.

Clustering purity is different for categorical data.

**Table 1.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=20%

| Clustering | Number of Clusters K | Distribution from Clusters | Purity of Clusters In % | SM=20% | | | |
|---|---|---|---|---|---|---|---|
| | | | | d= $U\ SM_i$ | CART | CHAID | NN |
| K-Means | 2 | 306/263 | 90.8 | 61+53 | 75.6 | 75.6 | 78.8 |
| K-Medoid | 2 | 318/251 | 91.8 | 64+50 | 75.2 | 75.3 | 78.4 |
| PAM | 2 | 345/224 | 89.4 | 69+45 | 74.3 | 72.4 | 76.6 |
| CLARA | 2 | 350/219 | 89.2 | 70+44 | 72.8 | 72.5 | 76.9 |

Breast cancer data has divided into clusters using three major clustering algorithms K-Means, K-Medoid, PAM and CLARA. 20% of each cluster instances taken as sample SM and merged them into single cluster denoted by "d". Classifiers CRT, CHAID and NN have been trained and tested for accuracy. NN gave the maximum accuracy 78.8 % under K-Means clustering algorithm when compared with others. The Fig is given below for comparison.

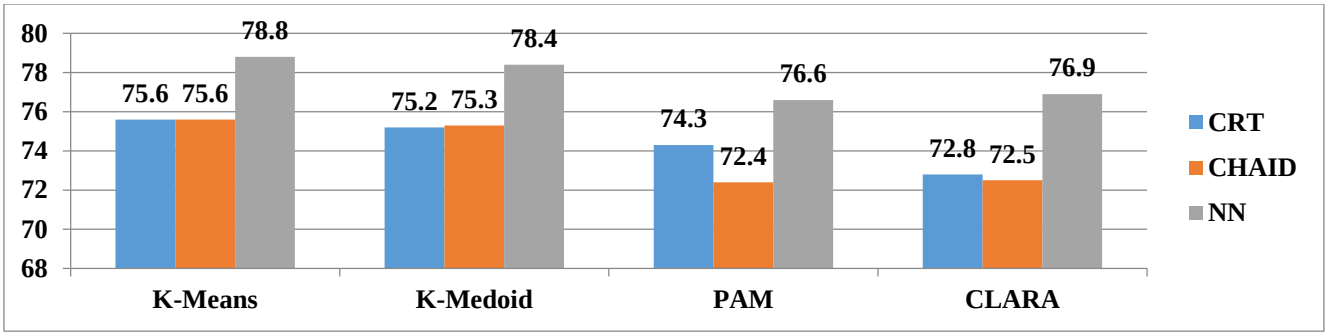


**Fig1.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=20%

**Table2.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=40%

| **Clustering** | **Number of Clusters K** | **Distribution from Clusters** | **Purity of Clusters In %** | **SM=40%** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **d= $U\ SM_i$** | **CART** | **CHAID** | **NN** |
| K-Means | 2 | 306/263 | 90.8 | 122/105 | 80.2 | 81.6 | 85.9 |
| K-Medoid | 2 | 318/251 | 91.8 | 127/100 | 81.6 | 82.9 | 86.4 |
| PAM | 2 | 345/224 | 89.4 | 138/90 | 79.6 | 80.3 | 84.3 |
| CLARA | 2 | 350/219 | 89.2 | 140/88 | 79.8 | 81.4 | 83.6 |

Breast cancer data has divided into clusters using four major clustering algorithms like above. Among all clusters 40% of instances taken as sample *SM* and merged them into single cluster denoted by "d". Classifiers CRT, CHAID and NN have been trained and tested for accuracy. NN gave the maximum accuracy 86.4 % under K-Medoid clustering algorithm

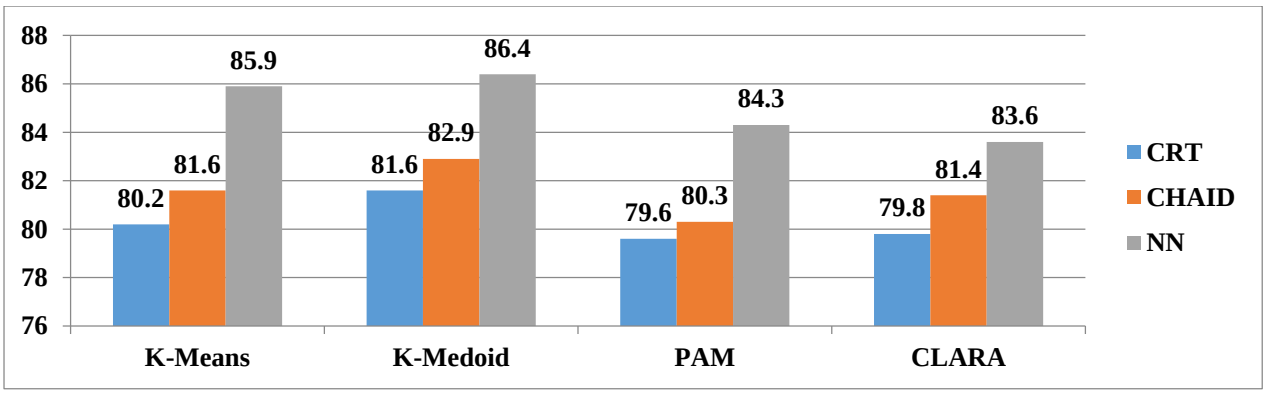


**Fig2.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=40%

**Table3.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=60%

| **Clustering** | **Number of Clusters K** | **Distribution from Clusters** | **Purity of Clusters In %** | **SM=60%** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **d= $U\ SM_i$** | **CART** | **CHAID** | **NN** |
| K-Means | 2 | 306/263 | 90.8 | 184/158 | 90.2 | 95.8 | 98.3 |
| K-Medoid | 2 | 318/251 | 91.8 | 191/151 | 90.5 | 96.2 | 98.8 |
| PAM | 2 | 345/224 | 89.4 | 207/134 | 92.6 | 94.6 | 96.4 |
| CLARA | 2 | 350/219 | 89.2 | 210/131 | 92.9 | 94.8 | 96.1 |

When 60% of sample taken from two clusters NN gave the maximum accuracy under the k-Medoid, this is the maximum accuracy when compared with 20%, 40% sample size from clusters. From all these experiments K-Means and K-Medoid algorithms suited well for small datasets.

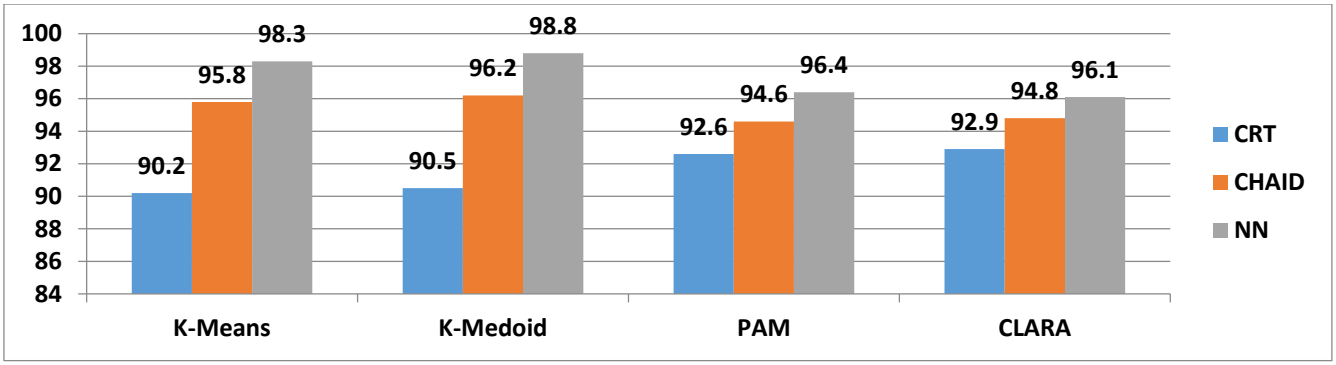


**Fig3.** Breast Cancer Data Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=60%

**Table4.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=20%

| Clustering | Number of Clusters K | Distribution from Clusters | Purity of Clusters In % | SM=20% | | | |
|---|---|---|---|---|---|---|---|
| | | | | d= $U\ SM_i$ | CART | CHAID | NN |
| K-Means | 2 | 510/570 | 58.8 | 102+114 | 70.6 | 70.1 | 74.2 |
| K-Medoid | 2 | 520/560 | 60.8 | 104+112 | 70.2 | 70.8 | 74.2 |
| PAM | 2 | 580/500 | 74.4 | 116+100 | 84.2 | 82.5 | 86.3 |
| CLARA | 2 | 564/516 | 76.8 | 113+103 | 82.6 | 82.3 | 86.4 |

Mice protein data is a large data with 1080 number of instances. Number of attributes is nearly 77. Similar process has been applied on Mice Protein data. When the data divided into two clusters and 20% of sample instances taken from two clusters d has 216 instances from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 86.4 under CLARA than other clustering algorithms.

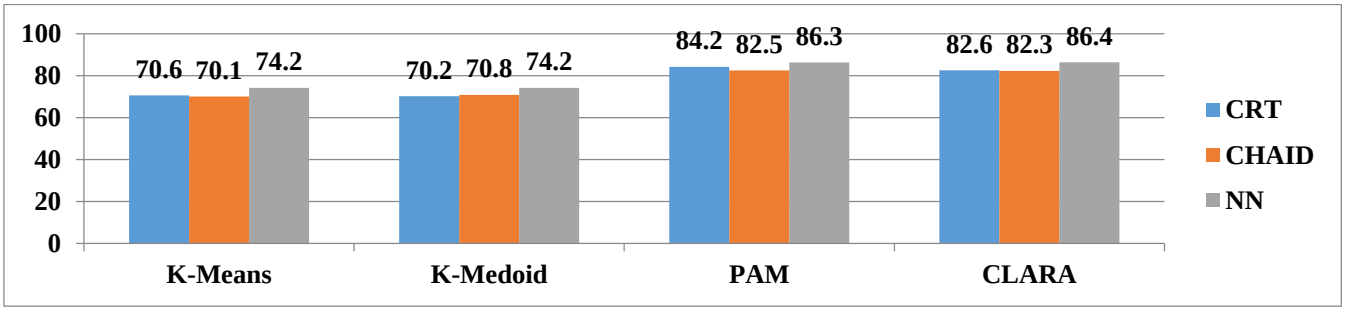


**Fig4.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=20%

**Table 5.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=40%

| Clustering | Number of Clusters K | Distribution from Clusters | Purity of Clusters In % | SM=40% | | | |
|---|---|---|---|---|---|---|---|
| | | | | d= $U\ SM_i$ | CART | CHAID | NN |
| K-Means | 2 | 510/570 | 58.8 | 204+228 | 71.4 | 71.4 | 75.8 |
| K-Medoid | 2 | 520/560 | 60.8 | 208+224 | 72.8 | 72.6 | 76.6 |
| PAM | 2 | 580/500 | 74.4 | 232+200 | 83.9 | 84.7 | 88.4 |
| CLARA | 2 | 564/516 | 76.8 | 226+206 | 84.8 | 84.8 | 88.9 |

When the data divided into two clusters and 40% of sample instances taken from two clusters d has 432 instances out of 1080 from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 88.9 under CLARA than other clustering algorithms.

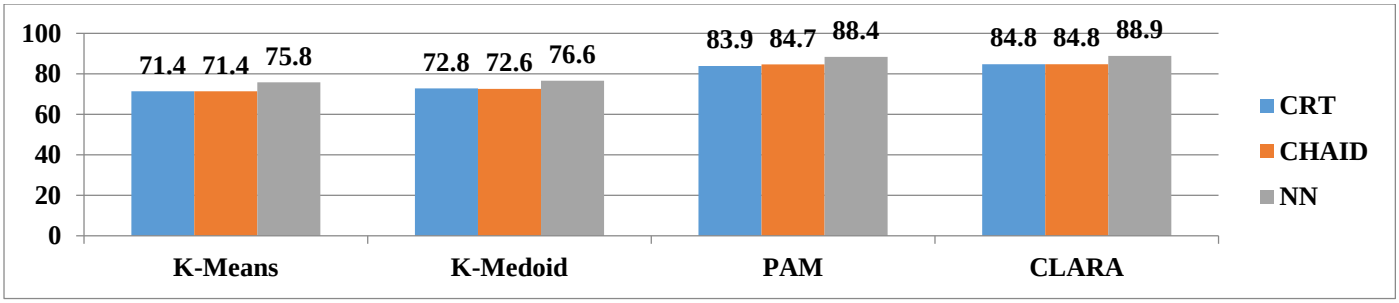


**Fig5.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=40%

**Table6.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=60%

| Clustering | Number of Clusters K | Distribution from Clusters | Purity of Clusters In % | SM=60% | | | |
|---|---|---|---|---|---|---|---|
| | | | | d= $U\ SM_i$ | CART | CHAID | NN |
| K-Means | 2 | 510/570 | 58.8 | 306+342 | 75.1 | 75.6 | 80.6 |
| K-Medoid | 2 | 520/560 | 60.8 | 312+336 | 78.6 | 79.4 | 81.9 |
| PAM | 2 | 580/500 | 74.4 | 348+300 | 87.2 | 87.9 | 95.8 |
| CLARA | 2 | 564/516 | 76.8 | 338+310 | 88.9 | 89.2 | 96.3 |

The same Procedure applied and divided into two clusters by taking 40% of sample instances from two clusters, d has 648 instances out of 1080 from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 96.3 under CLARA than other clustering algorithms.

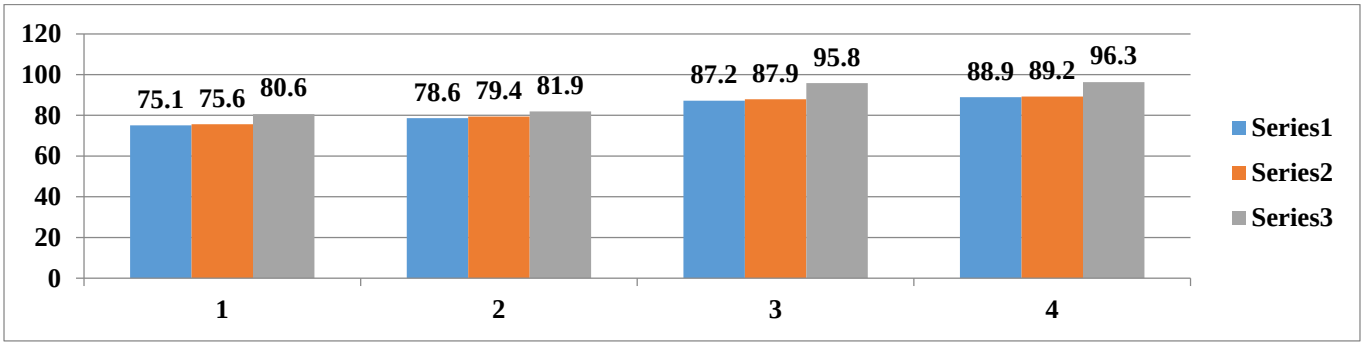


**Fig6.** Mice Protein Clustering with purities for K=2 and their Classifier Accuracies in percentages for SM=60%

**Table7.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=20%

| Clustering | Number of Clusters K | Distribution from Clusters | Purity of Clusters In % | SM=20% | | | |
|---|---|---|---|---|---|---|---|
| | | | | d= $U\ SM_i$ | CART | CHAID | NN |
| K-Means | 4 | 190/246/306/338 | 76.8 | 38+50+62+68 | 75.8 | 76.6 | 79.2 |
| K-Medoid | 4 | 196/242/303/339 | 78.8 | 39+49+61+68 | 75.4 | 76.4 | 79.2 |
| PAM | 4 | 150/204/309/417 | 75.6 | 30+41+62+84 | 86.6 | 84.8 | 86.8 |
| CLARA | 4 | 156/195/310/419 | 75.1 | 31+39+6284 | 84.4 | 85.4 | 86.2 |

When the data divided into four clusters and 20% of sample instances taken from four clusters d has 218 instances from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 86.8 under PAM than other clustering algorithms.

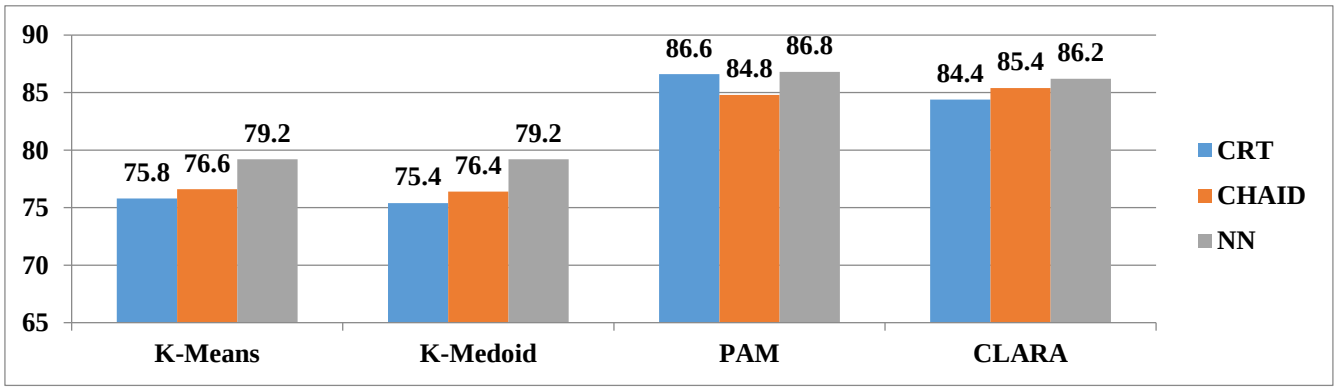


**Fig7.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=20%

**Table8.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=40%

| **Clustering** | **Number of Clusters K** | **Distribution from Clusters** | **Purity of Clusters In %** | **SM=40%** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **d=** $U\,SM_i$ | **CART** | **CHAID** | **NN** |
| K-Means | 4 | 190/246/306/338 | 76.8 | 76+98+122+135 | 79.2 | 79.1 | 82.7 |
| K-Medoid | 4 | 196/242/303/339 | 78.8 | 78+97+122+136 | 81.3 | 80.3 | 84.3 |
| PAM | 4 | 150/204/309/417 | 75.6 | 60+82+124+167 | 86.5 | 86.4 | 87.4 |
| CLARA | 4 | 156/195/310/419 | 75.1 | 62+78+124+168 | 86.9 | 86.2 | 89.6 |

When the data divided into four clusters and 40% of sample instances taken from four clusters d has 431 instances from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 89.6 under CLARA than other clustering algorithms.

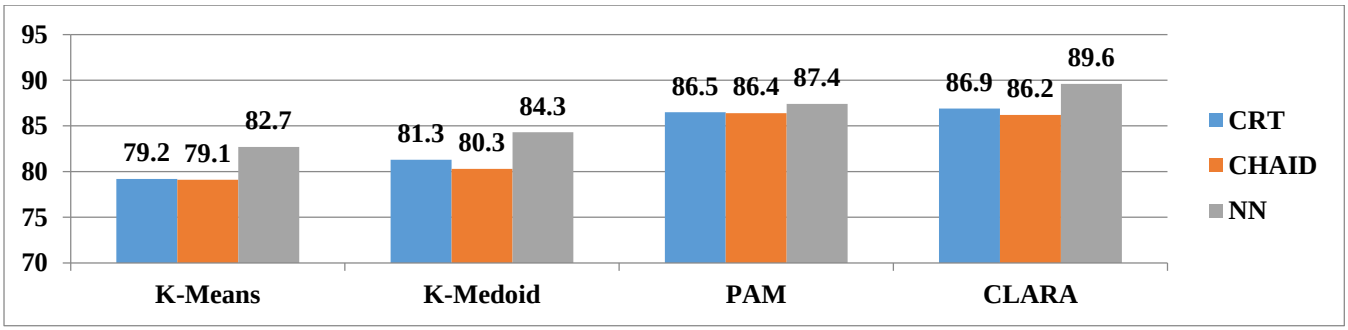


**Fig8.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=40%

Similar procedure applied to divide the data into four clusters and 40% of sample instances taken from four clusters d has 431 instances from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 89.6 under CLARA than other clustering algorithms.

**Table9.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=60%

| **Clustering** | **Number of Clusters K** | **Distribution from Clusters** | **Purity of Clusters In %** | **SM=60%** | | | |
|---|---|---|---|---|---|---|---|
| | | | | **d=** $U\, SM_i$ | **CART** | **CHAID** | **NN** |
| K-Means | 4 | 190/246/306/338 | 76.8 | 114+148+184+203 | 79.2 | 83.7 | 93.2 |
| K-Medoid | 4 | 196/242/303/339 | 78.8 | 118+145+182+203 | 81.3 | 84.5 | 95.2 |
| PAM | 4 | 150/204/309/417 | 75.6 | 90+122+185+250 | 88.5 | 87.4 | 98.6 |
| CLARA | 4 | 156/195/310/419 | 75.1 | 94+117+186+251 | 86.9 | 88.6 | 98.8 |

For 60% of sample instances taken from four clusters d has 649 instances from all clustering algorithms. CART, CHAID and NN have been trained and tested for accuracy. Accuracies are shown in the above table and in Fig below. NN gave good accuracy 98.8 under CLARA than other clustering algorithms.

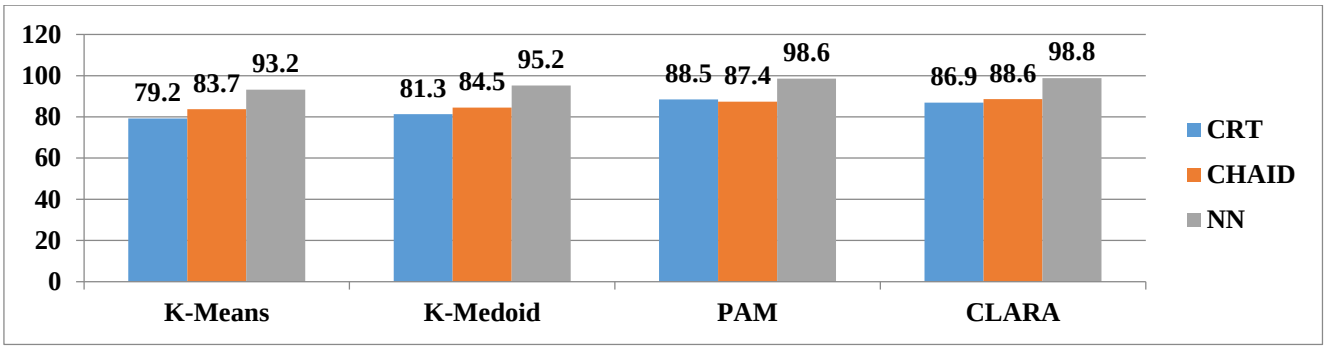


**Fig9.** Mice Protein Clustering with purities for K=4 and their Classifier Accuracies in percentages for SM=60%

**Table 10.** Acronyms

| Symbol | Abbreviation |
|---|---|
| U | Union |
| *Ci* | $i^{th}$ Cluster |
| *D* | Data Set |
| *SMi* | Sample % from $i^{th}$ cluster |
| *K* | Number of Clusters |
| PAM | Partition among Medoids |
| CLARA | Clustering Large Applications |
| *K-Means* | K-Means Clustering Algorithm |
| *K-Medoids* | K-Medoids Clustering Algorithm |
| CRT | Classification and Regression Tree |
| CHAID | Chi Square Automatic Interaction Detector |
| NN | Neural Network |

## VI. CONCLUSION & FUTURE WORK

When the classifiers applied directly on Breast Cancer data CRT gave 65.6% accuracy, CHAID gave 67.3% and NN gave 75.2% accuracy. For Mice Protein data CRT, CHAID and NN gave 74.2%, 76.5% and 85.3% respectively. We have achieved good accuracy by using Cluster sampling approach up to 98.6%. In future we can extend the same approach for categorical data as well as for mixed datasets.

## REFERENCES

[1] A. Hinneburg C. C. Aggarwal, , and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. Int. Conf. Database Theory, vol. 1, 2001, pp. 420–434.

[2] S. Abdallah, L. Du, and G. I. Webb, "Data preparation," in Encyclopedia of Machine Learning and Data Mining, C. Sammut and G. I. Webb,Eds.Boston,MA,USA:SpringerUS,2017,pp.318–327.

[3] L. Breiman, "Manual–setting up, using and understanding random forests v4. 0," 2003. [Online]. Available: https://www.stat. berkeley.edu/ breiman/Using_random_forests_v4.0.pdf

[4] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Amsterdam, The Netherlands: Elsevier, 2011.

[5] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," Appl. Stat., vol. 28, pp. 100–108, 1979.

[6] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program PAM)," in Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ, USA: Wiley, pp. 68–125, 1990.

[7] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining Knowl. Discovery, vol. 2, no. 3, pp. 283–304, 1998.

[8] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," in Proc. 15th Int. Conf. Data Eng., 1999, pp. 512–521.

[9] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 90–105, 2004.

[10] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1998, pp. 94–105

[11] .M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," J. Mach Learn Res., vol. 11, pp. 2487–2531, 2010.

[12] [N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "The role of hubness in clustering high-dimensional data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 739–751, Mar. 2014.

[13] E. Allen, S. Horvath, F. Tong, P. Kraft, E. Spiteri, A. D. Riggs, and Y. Marahrens, "High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes," Proc. Nat. Acad. Sci. USA, vol. 100, no. 17, pp. 9940–9945, 2003.

[14] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," J. Comput. Fig. Stat, vol. 15, no. 1, pp. 118–138, 2006.

[15] B. Azarnoush, J. M. Bekki, G. C. Runger, B. L. Bernstein, and R. K. Atkinson, "Toward a framework for learner segmentation," J. Educ. Data Mining, vol. 5, no. 2, pp. 102–126, 2013.

[16] Q. Zhang and I. Couloigner, "A new and efficient k-medoid algorithm for spatial clustering," Proc. Comput. Sci. Appl, 2005, pp. 207–224.

[17] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?" in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217–235.

[18] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," Expert Syst. Appl., vol. 36, no. 2, pp. 3336– 3341, 2009.

[19] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," Knowl.-Based Syst., vol. 30, pp. 129–135, 2012.

[20] [21] Z. He, X. Xu, and S. Deng, "Attribute value weighting in k-modes clustering," Expert Syst. Appl., vol. 38, pp. 15 365– 15 369, 2011.

[21] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[22] W. M. Rand, "Objective criteria for the evaluation of clustering methods," J. Am. Stat. Assoc., vol. 66, no. 336, pp. 846–850, 1971.

[23] H. Kremer, P. Kranen, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, "An effective evaluation measure for clustering on evolving data streams," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 868–876.

[24] G. Melli, "The datgen dataset generator. version 3.1," 1999. Available: http://www.datasetgenerator.com

[25] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[26] P. Franti, O. Virmajoki, and V. Hautamaki, "Fast agglomerative clustering using a k-nearest neighbor Fig," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 11, pp. 1875–1881, 2006.

[27] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R. News, vol. 2, no. 3, pp. 18–22, 2002, http:// CRAN.R-project.org/doc/Rnews/ [

[28] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, "Cluster: Cluster analysis basics and extensions," 2017.

[29] G. Szepannek, "R package 'clustmixtype'," 2017. [Online]. Available: https://CRAN.R-project.org/package=clustMixType

[30] N.JapkowiczandM.Shah,EvaluatingLearningAlgorithms:AClassificationPerspective.Cambridge,U.K.:CambridgeUniv.Press,2011.

[31] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," J. Mach Learn Res, vol. 7, pp. 1–30, 2006.