RESEARCH ARTICLE OPEN ACCESS

Cluster Ensemble Approach for Clustering Mixed Data

Honorine Mutazinda A[1],Mary Sowjanya [2],O.Mrudula[3]

[1,2,3](M.Tech, Department of Computer Science and Systems Engineering, Andhra University/College of Engineering, Visakhapatnam India)

----------**********----------

Abstract:

The paper presents a clustering ensemble method based on ensemble clustering approach for mixed data. A clustering ensemble is a paradigm that seeks to best combine the outputs of several clustering algorithms with a decision fusion function to achieve a more accurate and stable final output.

Most traditional clustering algorithms are limited to handling datasets that contain either numeric or categorical attribute and these algorithms were not generally scalable for large datasets. However; datasets with mixed types of attributes are common in real life data mining applications. So a novel divide-and-conquer technique is designed and implemented to solve this problem.

First, the original mixed dataset is divided into two sub-datasets: the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of datasets are employed to produce corresponding clusters. Last, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which the categorical data clustering algorithm is used to get the final clusters.

Keywords — **Clustering, Novel divide-and-conquer, Mixed Dataset, Numerical Data, and Categorical Data.**

----------**********----------

I. Introduction

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining models (prediction and description) are achieved by using the following primary data mining tasks: Classification, Regression, Clustering, Summarization, and Dependency modelling and Change and Deviation Detection. Clustering groups elements in a data set in accordance with its similarity such that elements in each cluster are similar while elements from different clusters are dissimilar. It involves analyzing or processing multivariate data, such as: characterizing customer groups based on purchasing patterns, categorizing Web documents, grouping genes and proteins that have similar functionality, grouping spatial locations prone to earthquakes based on seismological data, etc. Clustering ensembles or clustering fusion is the integration of results from various clustering algorithms using a consensus function to yield stable results. The idea of combining different clustering results (cluster ensemble or cluster aggregation) emerged as an alternative approach for improving the quality of the results of clustering algorithms.

In this paper a cluster ensemble approach using divide and conquer technique has been designed and implemented to deal with such type of mixed datasets. So, the initial dataset is divided into sub datasets namely numerical and categorical.
Then clustering algorithms designed for numerical and categorical datasets can be employed to produce corresponding clusters. Finally, the clustering results from the above step are combined as a categorical dataset on which the same categorical clustering algorithm or any other can be used to produce the final output clusters.

II. Related Work

A. Clustering Mixed Data

Most of the traditional clustering algorithms are designed to focus either on numeric data or on categorical data. The collected data in real world often contain both numeric and categorical attributes. It is difficult for applying traditional clustering algorithm directly into these kinds of data. Typically, when people need to apply traditional distance-based clustering algorithms to group these types of data, a numeric value will be assigned to each category in this attributes. Some categorical values, for example "low", "medium" and "high", can easily be transferred into numeric values. But if categorical attributes contain the values like "red", "white" and "blue" … etc., it cannot be ordered naturally.

Due to the differences in their features, in order to group these assorted data, it is good to exploit the clustering ensemble method which uses split and merge approach to solve this problem.For clustering mixed type attributes in [1] Ming-Yi Shih presented a new two-step clustering method is presented to find clusters on Mixed Categorical and Numeric Data. Items in categorical attributes are processed to construct the similarity or relationships among them based on the ideas of co-occurrence; then all categorical attributes can be converted into numeric attributes based on these constructed relationships. Finally, since all categorical data are converted into numeric, the existing clustering algorithms can be applied to the dataset without pain.

Jongwoo Lim1 , Jongeun Jun2 in [2] proposed a clustering framework that supports clustering of datasets with mixed attribute type (numerical, categorical), while minimizing information loss during clustering. They first utilize an entropy based measure of categorical attributes as a criterion function for similarity. Second, based on the results of entropy based similarity, they extract candidate cluster numbers and verify their weighting scheme with pre-clustering results. Finally, they cluster the mixed attribute type datasets with the extracted candidate cluster numbers and the weights.

Zhexue huang in[6] presented a k-prototypes algorithm which is based on the k-means paradigm but removes the numeric data limitation whilst preserving its efficiency. In the algorithm, objects are clustered against k prototypes. A method is developed to dynamically update the k prototypes in order to maximize the intra cluster similarity of objects. When applied to numeric data the algorithm is identical to the kmeans. To assist interpretation of clusters we use decision tree induction algorithms to create rules for clusters. These rules, together with other statistics about clusters, can assist data miners to understand and identify interesting clusters.

Jamil Al-Shaqsi and Wenjia Wang in [7] present a clustering ensemble method based on a novel three-staged clustering algorithm. Their ensemble is constructed with a proposed clustering algorithm as a core modeling method that is used to generate a series of clustering results with different conditions for a given dataset. Then, a decision aggregation mechanism such as voting is employed to find a combined partition of the different clusters. The voting mechanism considered only experimental results that produce intra-similarity value higher than the average intra-similarity value for a particular interval. The aim of this procedure is to find a clustering result that minimizes the number of disagreements between different clustering results.

B. Cluster Ensemble

Clustering fusion is the integration of results from various clustering algorithms using a consensus function to yield stable results.

The idea of combining different clustering results (cluster ensemble or cluster aggregation) emerged as an alternative approach for improving the quality of the results of clustering algorithms. It is based on the success of the combination of supervised classifiers. Given a set of objects, a cluster ensemble method consists of two principal steps: Generation, which is about the creation of a set of partitions of these objects, and Consensus Function, where a new partition, which is the integration of all partitions obtained in the generation step, is computed.

Generation Mechanisms:

Generation is the first step in clustering ensemble methods, in which the set of clusterings is generated and combined. It generates a collection of clustering solutions i.e., a cluster ensemble. Given a data set of n instances X = {X1,X2, · · · ,Xn} an ensemble constructor generates a cluster ensemble, represented as $\prod = \{\pi^1, \ldots, \pi^r\}$ *where r is the ensemble size (the number of clustering in the ensemble).*

Each clustering solution π^i is simply a partition of the data set X into Ki disjoint clusters of instances, represented as $\pi^i = c^i_1, \ldots c^i_k$.

It is very important to apply an appropriate generation process, because the ¯nal result will be conditioned by the initial clusterings obtained in this step.

In the generation step there are no constraints about how the partitions must be obtained. Therefore, in the generation process different clustering algorithms or the same algorithm with different parameters initialization can be applied.

1) ***Consensus Functions:*** The consensus function is the main step in any clustering ensemble algorithm. In this step, the final data partition or consensus partition P*, which is the result of any clustering ensemble algorithm, is obtained. However, the consensus among a set of clusterings is not obtained in the same way in all cases. There are two main consensus function approaches: objects co-occurrence and median partition.

III. A Cluster Ensemble Approach for Clustering Mixed Data

C. Overview

In This approach, instead of k means that assumes clusters are hyper-ellipsoidal and of similar sizes and which can't find clusters that vary in size to cluster numerical dataset, Chameleon an agglomerative hierarchical algorithm is chosen. This is due to the fact that Chameleon considers the internal characteristics of the clusters and can automatically adapt to the merged clusters. Also it can better model the degree of interconnectivity and closeness between each pair of clusters than K means. The existing Squeezer algorithm to cluster categorical data is retained as it is suitable for handling data streams and also can handle outliers effectively.

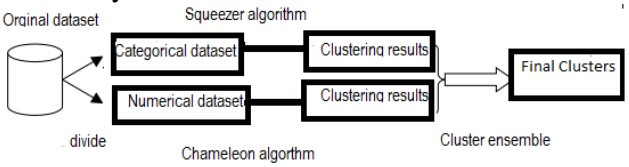


Figure 1.Overview of algorithm framework.

Algorithm

1. Splitting of the given data set into two parts. One for numerical data and another for categorical data.
2. Applying clustering Chameleon algorithms for numerical data set.
3. Applying clustering Squeezer algorithms for categorical data set.
4. Applying clustering Squeezer algorithms for categorical data set
5. Combining the output of step 2 and step 3 as cluster ensemble
6. Clustering the results using squeezer algorithm.

3) Final resultant clusters.Chameleon algorithm: Chameleon is a new agglomerative hierarchical clustering algorithm that overcomes the limitations of existing clustering algorithms. The Chameleon algorithm's key feature is that it accounts for both interconnectivity and closeness in identifying the most similar pair of clusters.

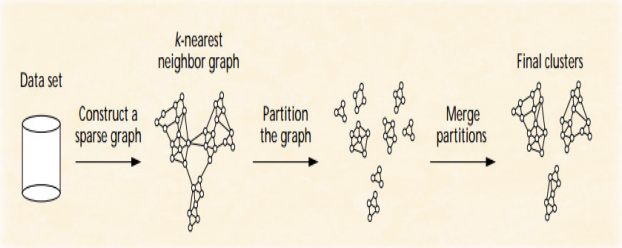


Figure 2.Overview of chameleon algorithm

Chameleon uses a two-phase algorithm, which first partitions the data items into sub-clusters and then repeatedly combines these sub-clusters to obtain the final clusters. It first clusters the data items into several sub-clusters that contain a sufficient number of items to allow dynamic modelling.

Chameleon uses a dynamic modelling framework to determine the similarity between pairs of clusters by looking at their relative interconnectivity (RI) and relative closeness (RC). Chameleon selects pairs to merge for which both RI and RC are high. That is, it selects clusters that are well interconnected as well as close together.

- **Relative interconnectivity**: Clustering algorithms typically measure the absolute interconnectivity between clusters Ci and Cj in terms of edge cut—the sum of the weight of the edges that straddle the two clusters, which we denote EC(Ci , Cj).

  Relative interconnectivity between clusters is their absolute interconnectivity normalized with respect to their internal interconnectivities. To get the cluster's internal interconnectivity, we sum the edges crossing a min-cut bisection that splits the cluster into two roughly equal parts.

Thus, the relative interconnectivity between a pair of clusters C_i and C_i is:

$$RI(C_i, C_j) = \frac{|EC(C_i, C_j)|}{\frac{|EC(C_i)| + |EC(C_j)|}{2}}$$

- **Relative closeness**: Relative closeness involves concepts that are analogous to those developed for relative interconnectivity.

The absolute closeness of clusters is the average weight (as opposed to the sum of weights for interconnectivity) of the edges that connect vertices in C_i to those in C_j.

To get a cluster's internal closeness, we take the average of the edge weights across a min-cut bisection that splits the cluster into two roughly equal parts. The relative closeness between a pair of clusters is the absolute closeness normalized with respect to the internal closeness of the two clusters:

$$RC(C_i, C_j) = \frac{\bar{S}EC(C_i, C_j)}{\frac{|C_i|}{|C_i| + |C_j|}\bar{S}EC(C_i) + \frac{|C_j|}{|C_i| + |C_j|}\bar{S}EC(C_j)}$$

Where $\bar{S}EC(C_i)$ and $\bar{S}EC(C_j)$ are the average weights of the edges that belong in the min-cut bisector of clusters C_i and C_j , and $SEC(C_i , C_j)$ is the average weight of the edges that connect vertices in C_i and C_j . Terms $|C_i|$ and $|C_j|$ are the number of data points in each cluster.

➢ **Advantages**

Existing clustering algorithms find clusters that fit some static model. Although effective in some cases, these algorithms can break down—that is, cluster the data incorrectly.

They break down when the data contains clusters of diverse shapes, densities, and sizes. Existing algorithms use a static model of the clusters and do not use information about the nature of individual clusters as they are merged.

***4)*Squeezer Algorithm**:

The Squeezer algorithm has n tuples as input and produces clusters as final results. Initially, the first tuple in the database is read in and a Cluster Structure (CS) is constructed with C = {1}. Then, the subsequent tuples are read iteratively. For each tuple, by our similarity function, we compute its similarities with all existing clusters, which are represented and embodied in the corresponding CSs. The largest value of similarity is selected out. If it is larger than the

given threshold, denoted as s, the tuple is put into the cluster that has the largest value of similarity. The CS is also updated with the new tuple. If the above condition does not hold, a new cluster must be created with this tuple. The algorithm continues until all tuples in the dataset are traversed.

The sub-function addNewClusterStructure() uses the new tuple to initialize Cluster and Summary, and then a new CS is created. The sub-function addTupleToCluster() updates the specified CS with new tuple. The subfunction simComputation(), which makes use of information stored in the CS to get the statistics based similarity.

Algorithm Squeezer(D,s) sim
Begin
1. While(D has unread tuple) {
2. Tuple=get Current Tuple(D)
3. If (tuple.tid==1) {
4. AddNewClusterStructure (tuple.tid)}
5. else {
6. for each existing cluster C
7. SimComputation(C,tuple)
8. get the max value of similarity: sim_max
9. Get the corresponding Cluster Index: index
10. if sim_max>= s
11. addTuppleToCluster(tuple,index)
12. else
13 addNewClusterStructure(tuple.tid)}
14. } handle outliers()
15. output ClusteringResults()
End

- **Advantages**

- The Squeezer algorithm only makes one scan over the dataset, thus, is highly efficient for disk resident datasets where the I/O cost becomes the bottleneck of efficiency.
- The algorithm is suitable for clustering data streams, where given a sequence of points, the objective is to maintain consistently good clustering of the sequence so far, using a small amount of memory and time.
- Outliers can be handled efficiently and directly.
- The algorithm does not require the number of desired clusters as an input parameter. This is very important for the user who usually does not know this number in advance.

IV. Methodology

A divide and conquer approach for mixed data using cluster ensemble works effectively either on pure numeric data or on pure categorical data.

In this approach first, the original mixed dataset is divided into two sub-datasets: the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of datasets can be employed to produce corresponding clusters. Here Chameleon algorithm is used for the numeric dataset where as Squeezer is used for the categorical dataset. In the Last step, the clustering results on the categorical and numeric dataset are combined as a categorical dataset, on which any categorical data clustering algorithm can be used to get the final clusters and Squeezer which was previously used for categorical data is again used for this purpose.

Algorithm

Step1. Splitting of the given data set into two parts. One for numerical data and another for categorical data.

Step 2. Applying clustering chameleon algorithms for numerical data set

Step 3. Applying clustering squeezer algorithms for categorical data set

Step 4. Combining the output of step 2 and step 3

Step 5. Clustering the results using squeezer algorithm.

Step6. Final cluster results.

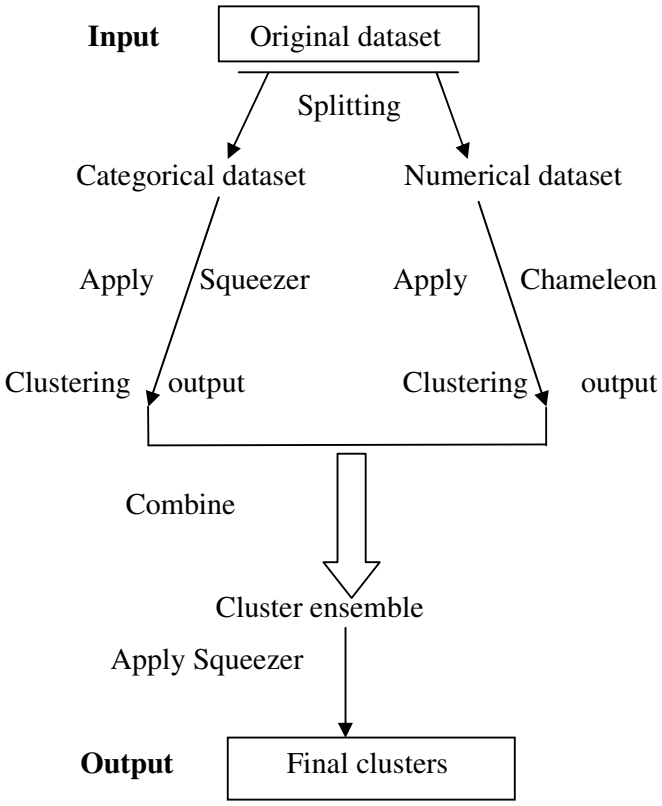
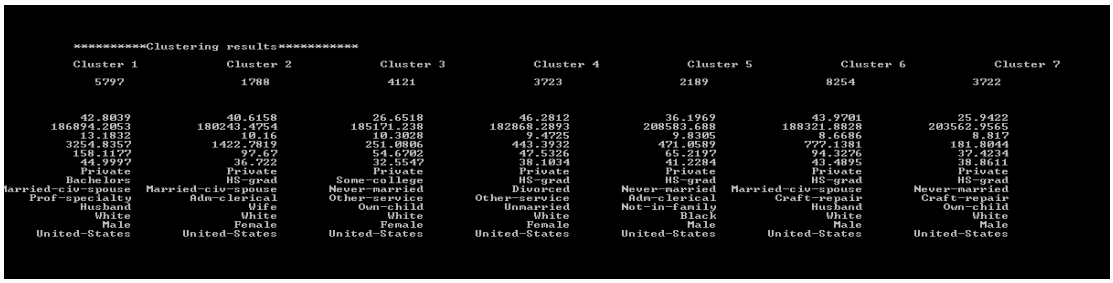


Figure 3.Algorithm for proposed framework

V. Experimental Results

D. Adult Dataset

Figure 4.Adult dataset



```
**********Clustering results**********
Cluster 1           Cluster 2           Cluster 3           Cluster 4           Cluster 5           Cluster 6           Cluster 7
5797                1788                4121                3723                2189                8254                3722

42.8039             40.6158             26.6518             46.2812             36.1969             43.9701             25.9422
186894.2053         180243.4754         185171.238          182868.2893         208583.688          188321.8828         203562.9565
13.1832             10.16               10.3028             9.4725              9.8305              8.6686              8.817
3254.8357           1422.7819           251.0806            443.3932            471.0589            777.1381            181.8044
158.1177            97.617              54.6742             47.5324             65.2197             94.3276             37.4234
44.9997             36.722              32.5547             38.1034             41.2284             43.4895             38.8611
Private             Private             Private             Private             Private             Private             Private
Bachelors           HS-grad             Some-college        HS-grad             HS-grad             HS-grad             HS-grad
Married-civ-spouse  Married-civ-spouse  Never-married       Divorced            Never-married       Married-civ-spouse  Never-married
Prof-specialty      Adm-clerical        Other-service       Other-service       Adm-clerical        Craft-repair        Craft-repair
Husband             Wife                Own-child           Unmarried           Not-in-family       Husband             Own-child
White               White               White               White               Black               White               White
Male                Female              Female              Female              Male                Male                Male
United-States       United-States       United-States       United-States       United-States       United-States       United-States
```

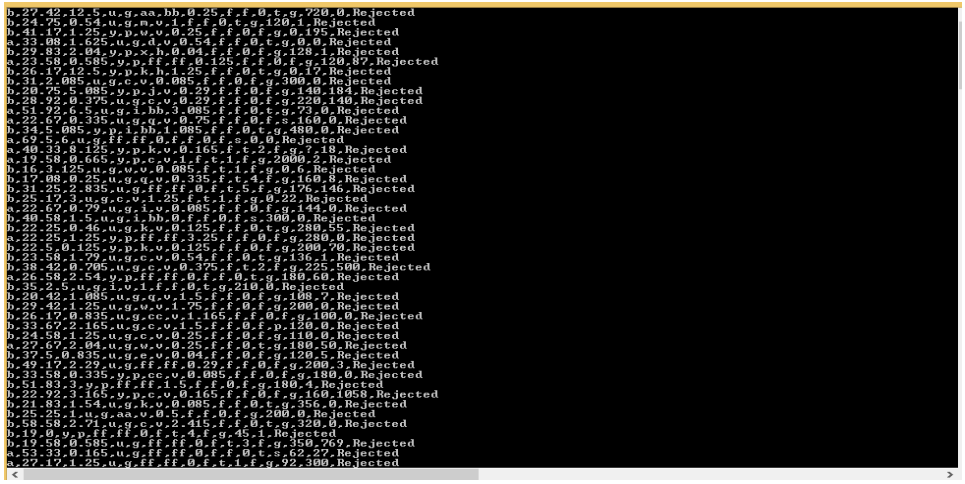
Figure 5.Clustering mixed data with k=7 on Adult dataset

E. Credit approval dataset

The credit approval dataset has 690 instances, each being described by 6 numeric and 9 categorical attributes. Instances were classified into two classes,"+" for approved label and "-"for rejected label.

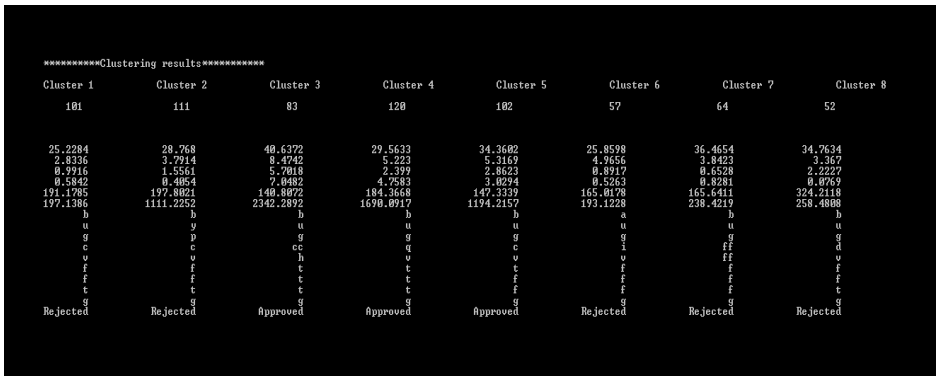
The comparison has been done with k-prototype algorithm that has been applied on credit approval dataset previously [10]. Clustering accuracy to measure clustering results is as follows:

$$r = \frac{\sum_{i=1}^{k} a_i}{n}$$



```
b,27.42,12.5,u,g,aa,bb,0.25,f,f,0,t,g,?,20,0,Rejected
b,24.75,0.54,u,g,m,v,1.5,f,f,0,t,g,120,1,Rejected
b,41.17,1.25,y,p,w,v,0.25,f,f,0,f,g,0,195,Rejected
a,33.08,1.625,u,g,d,v,0.54,f,f,0,t,g,0,0,Rejected
b,29.83,2.04,y,p,x,h,0.04,f,f,0,f,g,128,1,Rejected
a,23.58,0.585,y,p,ff,ff,0.125,f,f,0,f,g,120,87,Rejected
b,26.17,12.5,y,p,k,h,1.25,f,f,0,t,g,0,17,Rejected
b,31.2,085,u,g,c,v,0.085,f,f,0,f,g,300,0,Rejected
b,20.75,5.085,y,p,j,v,0.29,f,f,0,f,g,140,184,Rejected
b,28.92,0.375,u,g,c,v,0.29,f,f,0,f,g,220,140,Rejected
a,51.72,6.5,u,g,i,bb,3.085,f,f,0,t,g,73,0,Rejected
a,22.67,0.335,u,g,q,v,0.75,f,f,0,f,s,160,0,Rejected
b,34.5,085,y,p,i,bb,1.085,f,f,0,t,g,480,0,Rejected
a,69.5,6,u,g,ff,ff,0,f,f,0,f,s,0,0,Rejected
a,40.33,8.125,y,p,k,v,0.165,f,t,2,f,g,?,18,Rejected
a,19.58,0.665,y,p,c,v,1,f,t,1,f,g,2000,2,Rejected
b,16,3.125,u,g,w,v,0.085,f,t,1,f,g,0,6,Rejected
b,17.08,0.25,u,g,q,v,0.335,f,t,4,f,g,160,8,Rejected
b,31.25,2.835,u,g,ff,ff,0,f,t,5,f,g,176,146,Rejected
b,25.17,3,u,g,c,v,1.25,f,t,1,f,g,0,22,Rejected
a,22.67,0.79,u,g,i,v,0.085,f,f,0,f,g,144,0,Rejected
b,40.58,1.5,u,g,i,bb,0,f,f,0,f,s,300,0,Rejected
b,22.25,0.46,u,g,k,v,0.125,f,f,0,t,g,280,55,Rejected
a,22.25,1.25,y,p,ff,ff,3.25,f,f,0,f,g,280,0,Rejected
b,22.5,0.125,y,p,k,v,0.125,f,f,0,f,g,200,70,Rejected
b,23.58,1.79,u,g,c,v,0.54,f,f,0,t,g,136,1,Rejected
b,38.42,0.705,u,g,c,v,0.375,f,t,2,f,g,225,500,Rejected
a,26.58,2.54,y,p,ff,ff,0,f,f,0,t,g,180,60,Rejected
b,35.25,5,u,g,i,v,1.f,f,0,t,g,210,0,Rejected
b,20.42,1.085,u,g,q,v,1.5,f,f,0,f,g,100,7,Rejected
b,29.42,1.25,u,g,w,v,1.75,f,f,0,f,g,200,0,Rejected
b,26.17,0.835,u,g,cc,v,1.165,f,f,0,f,g,100,0,Rejected
b,33.67,2.165,u,g,c,v,1.5,f,f,0,f,p,120,0,Rejected
b,24.58,1.25,u,g,c,v,0.25,f,f,0,f,g,110,0,Rejected
a,27.67,2.04,u,g,w,v,0.25,f,f,0,t,g,180,50,Rejected
b,37.5,0.835,u,g,e,v,0.04,f,f,0,f,g,120,5,Rejected
b,49.17,2.29,u,g,ff,ff,0.29,f,f,0,f,g,200,3,Rejected
b,33.58,0.335,y,p,cc,v,0.085,f,f,0,f,g,180,0,Rejected
b,51.83,3,y,p,ff,ff,1.5,f,f,0,f,g,180,4,Rejected
b,22.92,3.165,y,p,c,v,0.165,f,f,0,f,g,160,1058,Rejected
b,21.83,1.54,u,g,k,v,0.085,f,f,0,t,g,356,0,Rejected
b,25.25,1,u,g,aa,v,0.5,f,f,0,f,g,200,0,Rejected
b,58.58,2.71,u,g,c,v,2.415,f,f,0,t,g,320,0,Rejected
b,19.0,y,p,ff,ff,0,f,t,4,f,g,45,1,Rejected
b,19.58,0.585,u,g,ff,ff,0,f,t,3,f,g,350,769,Rejected
a,53.33,0.165,u,g,ff,ff,0,f,f,0,t,s,62,27,Rejected
a,27.17,1.25,u,g,ff,ff,0,f,t,1,f,g,92,300,Rejected
```

Figure 6. Credit approval dataset



```
**********Clustering results**********
Cluster 1       Cluster 2       Cluster 3       Cluster 4       Cluster 5       Cluster 6       Cluster 7       Cluster 8
  101             111              83             120             102              57              64              52

 25.2284         28.768          40.6372         29.5633         34.3602         25.8598         36.4654         34.7634
  2.8336          3.7914          8.4742          5.223           5.3169          4.9656          3.8423          3.367
  0.9916          1.5561          5.7018          2.399           2.8623          0.8917          0.6528          2.2227
  0.5842          0.4054          7.0482          4.7583          3.0294          0.5263          0.8281          0.0769
191.1785        197.8021        149.8072        184.3668        147.3339        165.0178        165.6411        324.2118
197.1386        1111.2252       2342.2892       1690.0917       1194.2157       193.1228        238.4219        258.4008
       b               b               b               b               b               a               b               b
       u               y               u               u               u               u               u               u
       g               p               g               g               g               g               g               g
       c               c              cc               q               c               i              ff               d
       v               v               h               v               v               v              ff               v
       f               f               t               t               t               f               f               f
       f               f               t               t               f               f               f               f
       t               t               t               t               f               f               f               t
       g               g               g               g               g               g               g               g
Rejected        Rejected        Approved        Approved        Approved        Rejected        Rejected        Rejected
```

Figure 7. Clustering mixed data with k=8 on Credit Approval dataset

where n is the number of instances in the dataset,a1 is the number of instances occurring in both cluster I and its corresponding class, which has the maximum value.

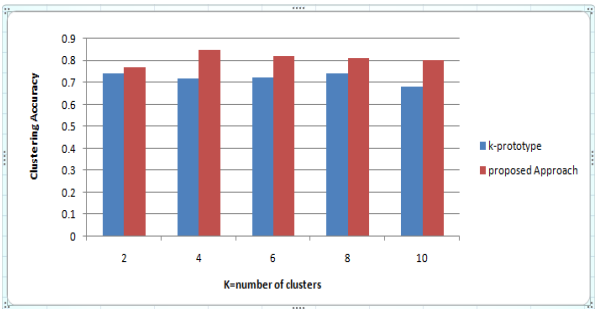


Figure 8. Clustering accuracy vs. number of clusters

VI. Conclusions

Cluster analysis has been widely a tool to various domains to discover the hidden and useful patterns inside datasets. Previous clustering algorithms mostly focus on either numeric or categorical, recently, approaches for clustering mixed attribute type datasets have been emerged, but they are mainly based on transforming categorical to numerical attributes. Such approaches have disadvantages of poor results due to the loss of information because important portion of attribute values can be dropped out while transformation. Therefore, the proposed framework approach divides the pure categorical dataset and the pure numeric dataset. Next, existing well established clustering algorithms designed for different types of datasets can be employed to produce corresponding clusters. Here Chameleon algorithm is used for the numeric dataset where as Squeezer is used for the categorical dataset. Clustering results on the categorical and numeric dataset are combined as a categorical dataset, which allows integration of other algorithms to produce corresponding clusters which leads to a better clustering accuracy.

In future work, other clustering algorithms for large scale dataset with mixed attribute types can be explored, also some weighting schemes on existing algorithms to perform well on their corresponding type of attributes to improve the proposed framework.

VII. References

[1] Ming-Yi Shih*, Jar-Wen Jheng and Lien-Fu Lai "*A Two-Step Method for Clustering Mixed Categroical and Numeric Data*" (2010).

[2] Jongwoo Lim , Jongeun Jun , Seon Ho Kim and Dennis McLeod "A Framework for Clustering Mixed Attribute Type Datasets ".

[3] M. V. Jagannatha Reddy1 and B. Kavitha" Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method "(2012).

[4] George Karypis Eui-Hong (Sam) Han Vipin Kumar"Chameleon:Hierarchical Clustering Using Dynamic Modeling".

[5] S. Anitha Elavarasi1 and J. Akilandeswari2 "survey on clustering algorithm and similarity measure for categorical data "(2014).
[6] zhexue huang "clustering large data sets with mixed numeric and categorical values".
[7] Jamil Al-Shaqsi and Wenjia Wang "A Clustering Ensemble Method for Clustering Mixed Data ".
[8] Zengyou He , Shengchun Deng "Squeezer: An efficient algorithm for clustering categorical data".
[9] Dileep Kumar Murala "Divide and Conquer Method for Clustering Mixed Numerical and Categorical Data"(2013).
[10] Zengyou He,Xiaofe I Xu, Shengehun Deng "Clustering Mixed Numeric and categorical Data: A cluster Ensemble Approach"
[11] divya d j1 & gayathri devi b2 "A meta clustering approach for ensemble problem ".
[12] Strehl and J. Ghosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions.Journal of Machine Learning Research, 3:583–617, 2002.
[13] G. Kharypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevelhypergraph partitioning: Applications in vlsi domain.In Proceedings of the Design and Automation Conference,1997.
[14] Topchy, A. K. Jain, and W. Punch. Combining multiple weak clusterings. In Proceedings IEEE International Conference on Data Mining, 2003.
[15] M. Meila. Comparing clusterings by the variation of information.In Proceedings of the Sixteenth Annual Conference of Computational Learning Theory (COLT), 2003.