Dynamic Dispatch Cluster Ensemble Approach for Mixed Attributes Dataset

Waale Angela Gboraloo [1], Chidiebere Ugwu[2]

[1]*Ken Saro Wiwa Polytechnic Bori, Department of Computer Science, Nigeria*
[2]*University of Port Harcourt, Department of Computer Science, Nigeria*

Abstract − *In recent time, data is growing binomially in almost all organizations in the world such as schools, hospitals, banks, which are usually of mixed attribute data values with numerical or categorical attribute data type. Several clustering systems with various clustering algorithms has been proposed to discover useful patterns that exist in such datasets, all adopting the same approach of splitting the dataset into two fragmented files and storing them on the storage device before subjecting them to clustering algorithms. This approach slows down the clustering process when there is large dataset. This paper presents a new dynamic dispatch cluster ensemble approach to clustering mixed attribute dataset based on ensemble technique where the attribute data type is automatically detected at run-time in place of outright splitting of the dataset into two subsets before clustering. The system utilized k means and Squeezer algorithms for clustering the various datasets. Object oriented design and Java programming language were used in the system development and implementation. The system was experimented on real life dataset obtained from UCL machine learning repository and results obtained were significantly different when compared to existing clustering systems. The process time was faster than the old systems because of the implicit and not explicit approach adopted in the system designs.*

Keywords − *Mixed Attributes Dataset, Clustering, Data Mining, Dynamic Dispatch and Cluster Ensemble.*

I. INTRODUCTION

Useful information that can aid management to make crucial business decisions may exist wasting in databases without cognate knowledge about it. These databases usually contain mixed dataset with numerical and categorical attribute values, such that the application of existing clustering system on these databases does not produce the expected result as they are specifically designed for either numerical or categorical attribute data values. Clustering is one of the data mining task that can be used to unveil hidden but useful information in a dataset. According to Singh [14], in clustering a given population of events or items can be partitioned or segmented into sets of similar elements.

Data mining involve the process of semi automatically analyzing large databases to find useful patterns. Data mining attempts to discover rules and patterns (knowledge) from data stored in a database [1].The knowledge discovered has several applications, the most commonly used is application that requires some sort of authentication and discovery [11].

Consequently,several authors have proposed the divide and conquer approach for clustering mixed attribute dataset, where these datasets are explicitly divided into two subsets – categorical and numerical and stored separately on the disk before applying existing algorithms to cluster them. These outright splitting of dataset into two subset leads to the following problems: reading is slow and applying recursive process to such fragmented files is complicated especially on billions of records with several attributes. many.Therefore, our major contribution in this paper is to present a dynamic dispatch cluster ensemble approach for clustering mixed attributes datasets.

Dynamic dispatch is the process of selecting which polymorphic function or method to call at run time. It is a prime characteristic of, object-oriented programming (OOP) languages and systems [7].

In imperative programming languages, the term "conditional statement" is usually used, whereas in functional programming, the terms "conditional expression" or "conditional construct" are preferred, because these terms all have distinct meanings.

Although dynamic dispatch is not usually classified as a conditional construct, it is another way to select between alternatives at runtime.

The instance mixed attribute data values are stored on the disk in Comma Separated Value (CSV) format or attribute related file format (ARFF). When the dataset is open in the system, the system automatically scanned for different attribute data value such as numerical and categoricalattribute value at runtime and dispatch appropriate algorithm to cluster each set in place of outright separation of dataset into two subsets and cluster them separately. The results from both attributes datatypes are combined as categorical data value using cluster ensemble and finally clustered with algorithm designed for categorical attribute dataset.

Cluster ensemble is technique that consolidate multiple clustering results into a single cluster often referred to as consensus solution. Cluster ensemble aims at generating a robust and stable clustering result compared to single clustering technique [18]. Clustering is a data mining technique that aims at discovering data colonies of interesting pattern in a dataset, the output produced by different algorithms is the assignment of data items to different colonies, identifying each colony with different cluster labels, items with same similarity are in the same cluster label and items with non-similarity are in another cluster label.

II. REVIEW OF RELATED WORK

Real world databases consist of mixed attribute dataset; mixed attribute dataset containsnumerical and categorical dataset.

Numerical attribute is an attribute whose data value can be quantified as integer or real and allow arithmetic operations to be perform on it while categorical attribute is an attribute whose data value cannot be quantified but can only be described and does not permit arithmetic operations. Clustering of these kind of dataset is a puzzling circumstance in data mining since most of the algorithms are purely for either numerical or categorical dataset.

Several systems with different algorithms has been proposed for clustering mixed attributes dataset, these systems explicitly separates the instance dataset into two different sets as seen in [15] where in the first stage the instance dataset was manually divided into two subsets, each with a pure kind of attribute datatype; categorical or numerical, in the second stage, each subset is clustered with its corresponding algorithm, the results are merged to form a categorical attribute datatype and in the third stage merged categorical attribute datatype is finally clustered with categorical clustering algorithm.

In cluster ensemble approach by Honorine *et al.*[3] the dataset was manually divided into two subsets: numerical and categorical dataset, cluster the datasets with Chameleon and Squeezer algorithm respectively, combined results as categorical and finally cluster it with Squeezer algorithm. The same divide and conquer approach was used in Asadi *et al.*[2] and Kavitha & Reddy[10] to cluster mixed attribute dataset, here the instance dataset was also divided into two subsets; numerical and categorical, cluster both numerical and categorical data set using Similarity Weight, combined both clusters as categorical and finally cluster it with Filter method.

In Sugana & Selvi[16] a fuzzy clustering method for mixed dataset is also proposed, where the dataset is divided into two sub datasets explicitly; pure numerical and pure categorical dataset and then applied fuzzy c-means clustering algorithm on numerical dataset and fuzzy c-mode on categorical dataset to generate clusters.

A Two-step method for clustering mixed attribute dataset where the instance dataset is read as initial input was proposed. In the first step the k means algorithm is applied to the dataset to divide it into different subset. Assign zero and one to numeric attribute values, convert categorical attribute as numeric.In step two, the based attribute is defined by selecting the attribute with the most number of item's, then count the frequency of co-occurrence among categorical. Then apply k means algorithm [8].

Zengyou *et al.*[17] proposed cluster ensemble based approach where the original dataset is divided into two subsets; numerical and categorical and appliedappropriate algorithm designed for each. The clustered results from both datasets where combined as categorical dataset using ensemble function and further cluster with squeezer algorithm.

Kprototypealgorithm [4] is a combination of k means and k mode algorithms to cluster mixed numeric and categorical values. K prototype algorithms defines a dissimilarity measure that take into account both numeric and categorical attributes by utilizing weighted sum of Euclidean distance.

Chameleon algorithm by Karypis[5]and cited in Rafsanjami *et al.*[9] is a hierarchical agglomerative clustering algorithm that operates in two phases. In the first phase, the dataset is partitions into sub-clusters using graph partitioning method and in the second phase it repeatedly merges the sub-clusters from the first stage to generate the final clusters. This algorithm has been proven to work well on clusters of different shapes, sizes and densities and also capable of handling huge dataset but has a worst-case of time complexity of O(n2).

III ANALYSIS OF EXISTING SYSTEMS

The analysis of the existing clustering system for clustering mixed dataset reveal that:

1. The instance mixed dataset was manually (explicitly) divided into two subsets; one set as numerical and the other set as categorical dataset and stored on the disk separately.
2. Cluster the categorical dataset using categorical clustering algorithm such as squeezer, rock etc.
3. Cluster the numerical dataset using numerical clustering algorithm such as chameleon, cure etc.
4. Combine the clusters from step 2 with step 3 as categorical dataset using ensemble approach
5. Cluster the final dataset using algorithms designed for categorical dataset.

The architecture of existing clustering system is illustrated in Figure 1 where divide and conquer method is used in clustering mixed attribute dataset.

Divide and conquer method of solving problems in computer help to optimize human effort in tackling complex problems but the explicit division of the instance dataset into two subsets and store it on the disk may be time consuming where the attributes are so many.

Therefore, in this paper we present a new automatic method to divide and conquer the problem using Dynamic Dispatch Cluster ensemble approach to cluster mixed numerical and categorical attributes.

IVDYNAMIC DISPATCH CLUSTER ENSEMBLE APPROACH

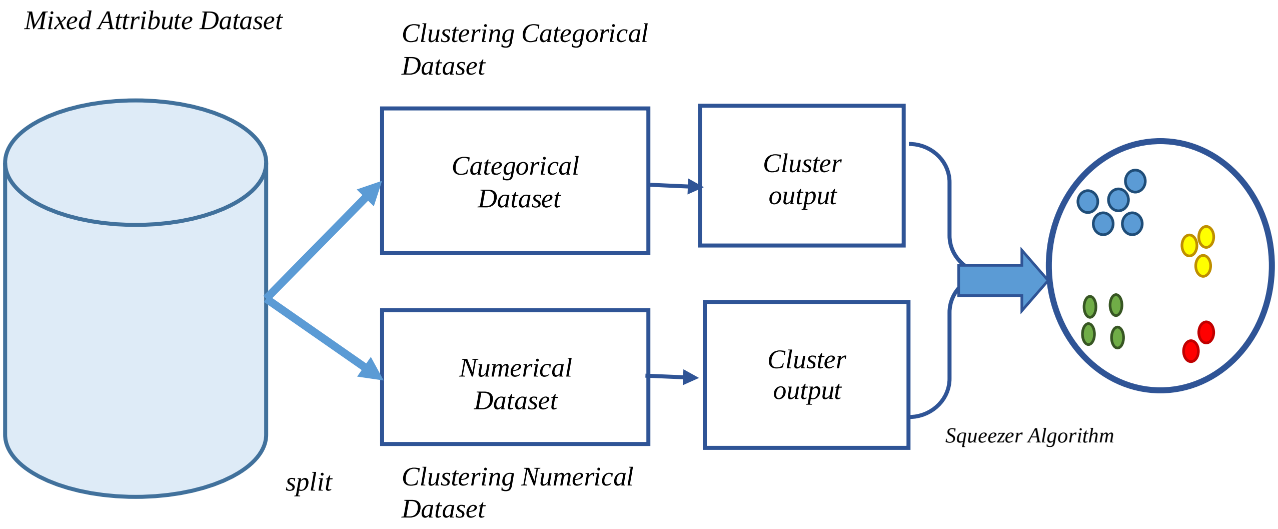


Figure 1: Architecture of existing system

Dynamic dispatch cluster ensemble approach is illustrated in Figure 2. In this approach, the instance dataset is read into the system, the system checks the attributes within the dataset then split it into two sub datasets implicitly as numerical and categorical dataset instead of splitting the dataset explicitly as observed in the existing system thereby optimizing the time and efficiency of the algorithm. If the dataset is numerical then cluster the subset with k means algorithm and if the subset is categorical then cluster the dataset with Squeezer algorithm, combined both results as categorical data value and cluster the result with squeezer algorithm to yield the final clusters.

The system retains squeezer algorithm for clustering categorical attribute dataset due its efficiency, scalability and capability to handle high dimensional data effectively in one scan thereby optimizing memory usage. While k means algorithm is used in place of Chameleon algorithm used in existing system for clustering numerical dataset due its efficiency to handle large dataset.

V. ALGORITHM FOR THE NEW SYSTEM

```
Input Dataset = Z
Attribute  =  S
If Z = MixedDataset
{
     {
        If  S = Numerical attributes
// Cluster numerical datasets
        Call kmeans module
else
// Cluster categorical datasets Call Squeezer
        module
end if
}

// Combine Clusters
Call ensemble module

// Final Clustering
Call Squeezer module

Display Output
}
```

A. K Means Algorithm

K means algorithm [6] is a popular algorithm that has been proven efficient by several authors and widely used for clustering numerical dataset despite it shortcoming like the choice of initial number of clusters which determine the final output of clusters. The steps involved:

1. Determine the number of cluster k randomly and assume the centroid or center of these clusters.

2. Compute the distance of each object to the centroids using Euclidean distance given as:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^{2}}$$

3. Group the data based on minimum distance (i.e. find the nearest centroid)

4. Repeat steps 2 and 3 until computation is stable

B. Squeezer Algorithm

Squeezer algorithm [18] is one of the algorithms designed for clustering categorical dataset which has been widely used by several authors because of the following features. Squeezer has been proven to be highly effective in handling large volume of data, it makes one scan over the dataset there by optimizing the input and output cost of memory usage, it is scalable and capable of producing high quality result.

The steps involve in squeezer include:

Input: Dataset (D) and Threshold (s)

1. Read the first tuple.
2. Generate the Cluster Structure (CS).
3. Read the next tuple and computes its similarity using support measure given as:

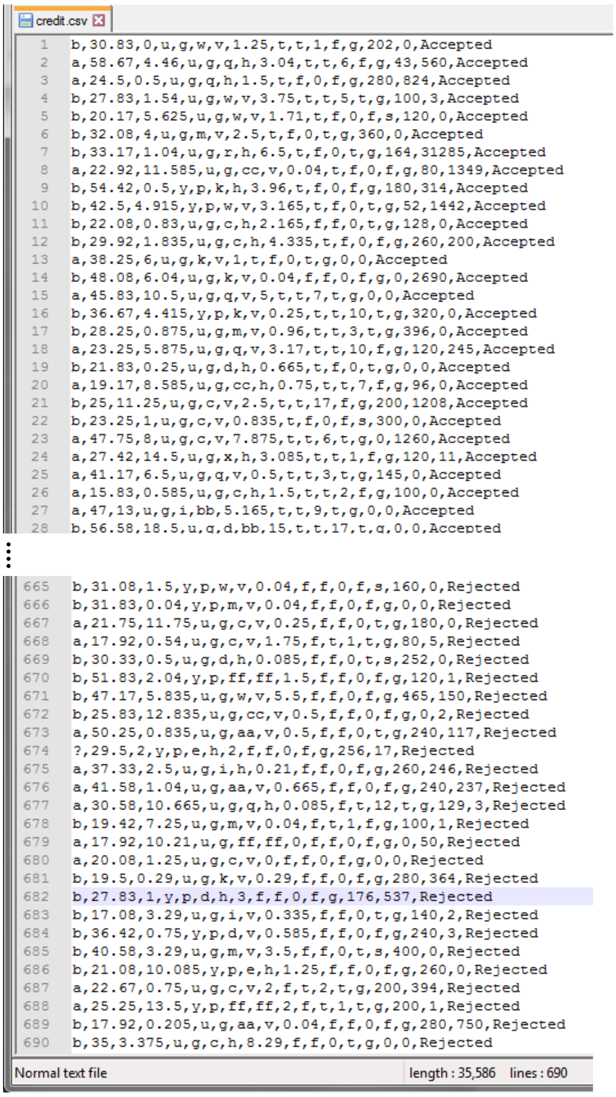
$$Sim(C, tid) = \sum_{i=1}^{m} \frac{\sup(a_i)}{\sum_j \sup(a_i)}$$

4. If the similarity is greater than the threshold "s". Add tuple to the existing Cluster Structure. Else assign tuple to the new Cluster Structure.
5. Repeat Step 2 through 4 until the end of the tuple.

Output: Cluster output.

VI. EXPERIMENTS AND RESULTS

Credit approval dataset from UCL machine learning repository was used to test the system. Credit approval dataset consist of 690 instances with 6 numericalattributes and 9 categorical (nominal) attributes, a sample view of the dataset is shown in figure 4. The classes consist of two groups: accepted (+) and rejected (-).

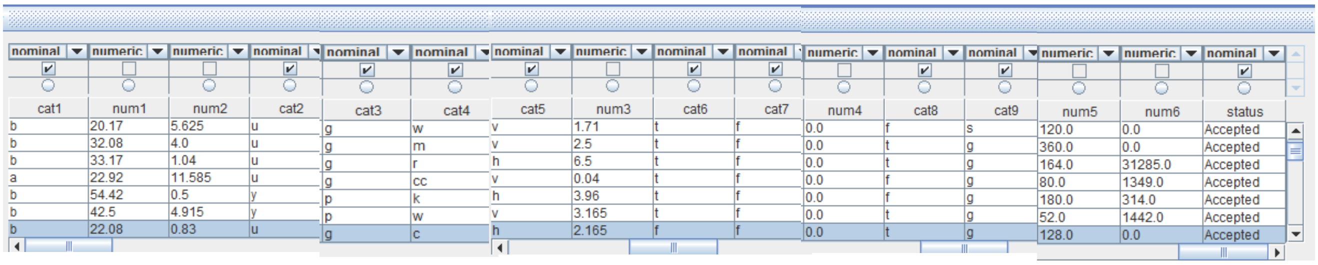


credit.csv

```
b,30.83,0,u,g,w,v,1.25,t,t,1,f,g,202,0,Accepted
a,58.67,4.46,u,g,q,h,3.04,t,t,6,f,g,43,560,Accepted
a,24.5,0.5,u,g,q,h,1.5,t,f,0,f,g,280,824,Accepted
b,27.83,1.54,u,g,w,v,3.75,t,t,5,t,g,100,3,Accepted
b,20.17,5.625,u,g,w,v,1.71,t,f,0,f,s,120,0,Accepted
b,32.08,4,u,g,m,v,2.5,t,f,0,t,g,360,0,Accepted
b,33.17,1.04,u,g,r,h,6.5,t,f,0,t,g,164,31285,Accepted
a,22.92,11.585,u,g,cc,v,0.04,t,f,0,f,g,80,1349,Accepted
b,54.42,0.5,y,p,k,h,3.96,t,f,0,f,g,180,314,Accepted
b,42.5,4.915,y,p,w,v,3.165,t,f,0,t,g,52,1442,Accepted
b,22.08,0.83,u,g,c,h,2.165,f,f,0,t,g,128,0,Accepted
b,29.92,1.835,u,g,c,h,4.335,t,f,0,f,g,260,200,Accepted
a,38.25,6,u,g,k,v,1,t,f,0,t,g,0,0,Accepted
b,48.08,6.04,u,g,k,v,0.04,f,f,0,f,g,0,2690,Accepted
a,45.83,10.5,u,g,q,v,5,t,t,7,t,g,0,0,Accepted
b,36.67,4.415,y,p,k,v,0.25,t,t,10,t,g,320,0,Accepted
b,28.25,0.875,u,g,m,v,0.96,t,t,3,t,g,396,0,Accepted
a,23.25,5.875,u,g,q,v,3.17,t,t,10,f,g,120,245,Accepted
b,21.83,0.25,u,g,d,h,0.665,t,f,0,t,g,0,0,Accepted
a,19.17,8.585,u,g,cc,h,0.75,t,t,7,f,g,96,0,Accepted
b,25,11.25,u,g,c,v,2.5,t,t,17,f,g,200,1208,Accepted
b,23.25,1,u,g,c,v,0.835,t,f,0,f,s,300,0,Accepted
a,47.75,8,u,g,c,v,7.875,t,t,6,t,g,0,1260,Accepted
a,27.42,14.5,u,g,x,h,3.085,t,t,1,f,g,120,11,Accepted
a,41.17,6.5,u,g,q,v,0.5,t,t,3,t,g,145,0,Accepted
a,15.83,0.585,u,g,c,h,1.5,t,t,2,f,g,100,0,Accepted
a,47,13,u,g,i,bb,5.165,t,t,9,t,g,0,0,Accepted
b,56.58,18.5,u,g,d,bb,15,t,t,17,t,g,0,0,Accepted
⋮
b,31.08,1.5,y,p,w,v,0.04,f,f,0,f,s,160,0,Rejected
b,31.83,0.04,y,p,m,v,0.04,f,f,0,f,g,0,0,Rejected
a,21.75,11.75,u,g,c,v,0.25,f,f,0,t,g,180,0,Rejected
a,17.92,0.54,u,g,c,v,1.75,f,t,1,t,g,80,5,Rejected
b,30.33,0.5,u,g,d,h,0.085,f,f,0,t,s,252,0,Rejected
b,51.83,2.04,y,p,ff,ff,1.5,f,f,0,f,g,120,1,Rejected
b,47.17,5.835,u,g,w,v,5.5,f,f,0,f,g,465,150,Rejected
b,25.83,12.835,u,g,cc,v,0.5,f,f,0,f,g,0,2,Rejected
a,50.25,0.835,u,g,aa,v,0.5,f,f,0,t,g,240,117,Rejected
?,29.5,2,y,p,e,h,2,f,f,0,f,g,256,17,Rejected
a,37.33,2.5,u,g,i,h,0.21,f,f,0,f,g,260,246,Rejected
a,41.58,1.04,u,g,aa,v,0.665,f,f,0,f,g,240,237,Rejected
a,30.58,10.665,u,g,q,h,0.085,f,t,12,t,g,129,3,Rejected
b,19.42,7.25,u,g,m,v,0.04,f,t,1,f,g,100,1,Rejected
a,17.92,10.21,u,g,ff,ff,0,f,f,0,f,g,0,50,Rejected
a,20.08,1.25,u,g,c,v,0,f,f,0,f,g,0,0,Rejected
b,19.5,0.29,u,g,k,v,0.29,f,f,0,f,g,280,364,Rejected
b,27.83,1,y,p,d,h,3,f,f,0,f,g,176,537,Rejected
b,17.08,3.29,u,g,i,v,0.335,f,f,0,t,g,140,2,Rejected
b,36.42,0.75,y,p,d,v,0.585,f,f,0,f,g,240,3,Rejected
b,40.58,3.29,u,g,m,v,3.5,f,f,0,t,s,400,0,Rejected
b,21.08,10.085,y,p,e,h,1.25,f,f,0,f,g,260,0,Rejected
a,22.67,0.75,u,g,c,v,2,f,t,2,t,g,200,394,Rejected
a,25.25,13.5,y,p,ff,ff,2,f,t,1,t,g,200,1,Rejected
b,17.92,0.205,u,g,aa,v,0.04,f,f,0,f,g,280,750,Rejected
b,35,3.375,u,g,c,h,8.29,f,f,0,t,g,0,0,Rejected
```

Normal text file   length : 35,586   lines : 690

Figure 3: Credit Approval dataset



nominal	numeric	numeric	nominal	nominal	nominal	nominal	numeric	nominal	nominal	numeric	nominal	nominal	numeric	numeric	nominal
☑	□	□	☑	☑	☑	☑	□	☑	☑	□	☑	☑	□	□	☑
○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
cat1	num1	num2	cat2	cat3	cat4	cat5	num3	cat6	cat7	num4	cat8	cat9	num5	num6	status
b	20.17	5.625	u	g	w	v	1.71	t	f	0.0	f	s	120.0	0.0	Accepted
b	32.08	4.0	u	g	m	v	2.5	t	f	0.0	t	g	360.0	0.0	Accepted
b	33.17	1.04	u	g	r	h	6.5	t	f	0.0	t	g	164.0	31285.0	Accepted
a	22.92	11.585	u	g	cc	v	0.04	t	f	0.0	f	g	80.0	1349.0	Accepted
b	54.42	0.5	y	p	k	h	3.96	t	f	0.0	f	g	180.0	314.0	Accepted
b	42.5	4.915	y	p	w	v	3.165	t	f	0.0	t	g	52.0	1442.0	Accepted
b	22.08	0.83	u	g	c	h	2.165	f	f	0.0	t	g	128.0	0.0	Accepted

Figure 4: Instance dataset automatically detected as Numerical and Categorical (Nominal)

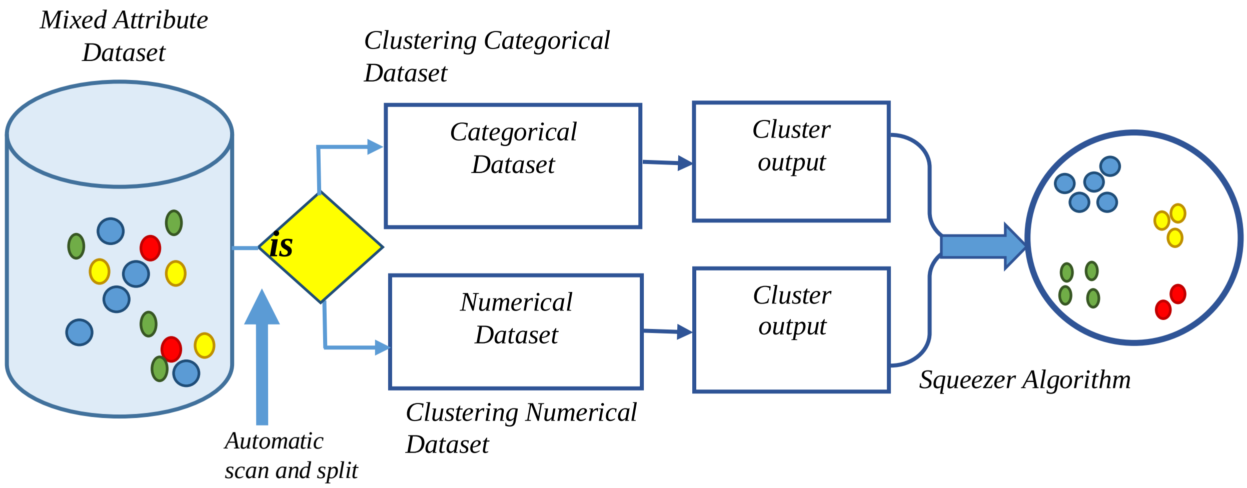
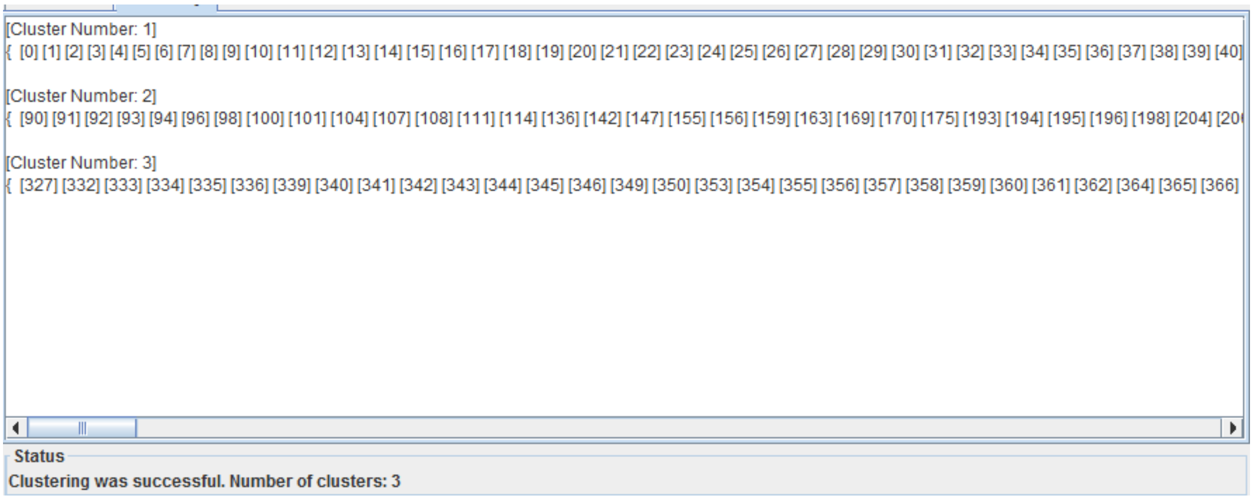


Figure 2: Architecture of New Dynamic Dispatch Cluster Ensemble approach

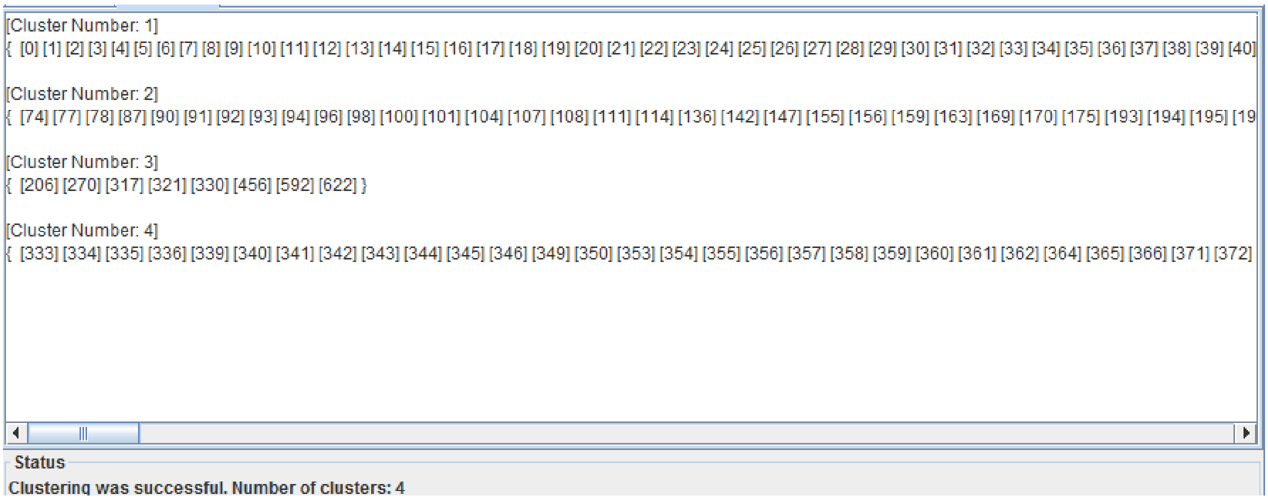


```
[Cluster Number: 1]
{  [0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40]

[Cluster Number: 2]
{  [90] [91] [92] [93] [94] [96] [98] [100] [101] [104] [107] [108] [111] [114] [136] [142] [147] [155] [156] [159] [163] [169] [170] [175] [193] [194] [195] [196] [198] [204] [20

[Cluster Number: 3]
{  [327] [332] [333] [334] [335] [336] [339] [340] [341] [342] [343] [344] [345] [346] [349] [350] [353] [354] [355] [356] [357] [358] [359] [360] [361] [362] [364] [365] [366]
```

Status

Clustering was successful. Number of clusters: 3

Figure 5: Clustering Result with k = 3



```
[Cluster Number: 1]
{  [0] [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40]

[Cluster Number: 2]
{  [74] [77] [78] [87] [90] [91] [92] [93] [94] [96] [98] [100] [101] [104] [107] [108] [111] [114] [136] [142] [147] [155] [156] [159] [163] [169] [170] [175] [193] [194] [195] [19

[Cluster Number: 3]
{  [206] [270] [317] [321] [330] [456] [592] [622] }

[Cluster Number: 4]
{  [333] [334] [335] [336] [339] [340] [341] [342] [343] [344] [345] [346] [349] [350] [353] [354] [355] [356] [357] [358] [359] [360] [361] [362] [364] [365] [366] [371] [372]
```

Status

Clustering was successful. Number of clusters: 4

Figure 6: Clustering Result with k = 4

When the dataset was feed into the system, the system automatically scanned and detected the attributes datatype of the dataset as numerical or categorical and checked the categorical as nominal while numerical is unchecked. These automatic-scan and detection of the attributes datatype is shown in the screen shot of figure 4. Figures 5 and 6 shows the clustering result of k = 3 and k = 4 respectively with the same dataset.

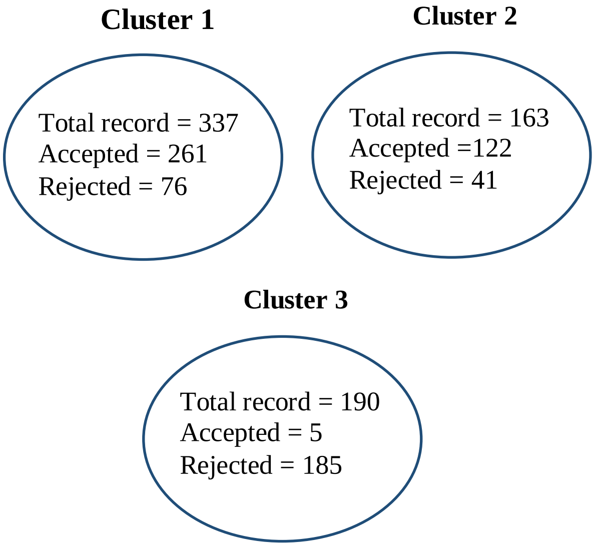
The value of k for number of clusters was randomly chosen by the system and cluster result are displayed with tuple index numbers. Certain clusters repeated in each runs showing the degree of similarity among the dataset. For example, cluster number 1 in figure 5 repeated in figure 6.

Result obtained from the system were compared with result obtained from other systems. It is significant to note that our system performs better than system with k prototype algorithm and all instances in the dataset were clustered, unlike system with *K*prototype algorithm that do not handle numerical attribute with missing values.

Figure 7 shows the chart of clustering accuracy against number of clusters.The Cluster accuracy (r) was computed using the formula given as:

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

where *n*is the totalnumber of instances in the dataset, a_i is the maximum number of class in a cluster *k*, with i = 1, 2, … n, for example the cluster accuracy of Figure 5 was computed as:



Cluster 1

Total record = 337
Accepted = 261
Rejected = 76

Cluster 2

Total record = 163
Accepted =122
Rejected = 41

Cluster 3

Total record = 190
Accepted = 5
Rejected = 185

$$r = \frac{261 + 122 + 185}{690}$$

$$= 0.823 \quad \approx 82\%$$

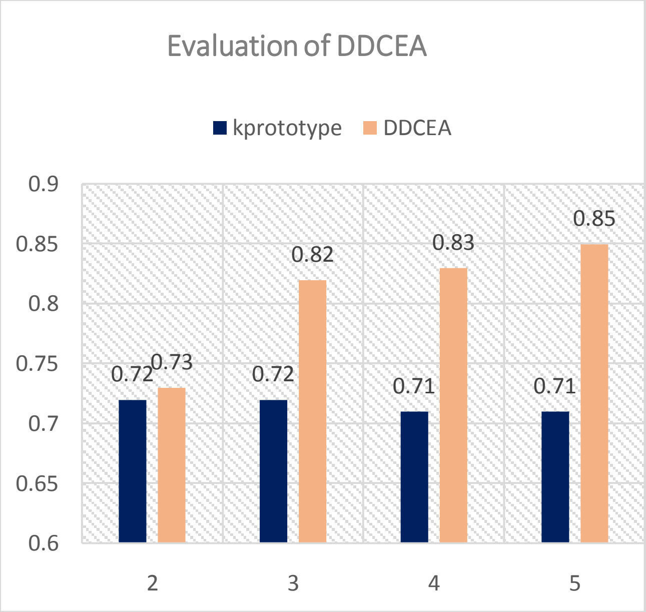


Figure 7: Cluster accuracy and Number of Clusters

The computed cluster accuracy was evaluatedwith system that runs k-prototype algorithm against credit approval dataset with a scale value of 0.5 starting at 0.60 to 0.90. With cluster number set to 2, both system perform almost the same with clustering accuracy of 0.72:0.73 for *K* prototype and DDCEA respectively as shown in the chart of Figure 7.

Clustering accuracy of our system was significantly higher than system with k prototype algorithm as cluster numbers increases from 2 to 3, 4 and 5 with cluster accuracies of 0.72:0.82%, 0.71:0.83% and 0.71:0.85% respectively.

VII CONCLUSION

There exist several systems for clustering mixed numerical and categorical attribute dataset. These systems adopt explicit division of mixed attribute dataset into two separate sets; numerical and categorical which may be time consuming and slow for recursive process where the attributes are too many to handle. Consequently, in this paper we present a new dynamic dispatched cluster ensemble approached to split the dataset implicitly into two separate sets; numerical and categorical before clustering. In this approach the dataset is read and scanned automatically to detect it as either numerical and categorical attributes dataset and then clustered accordingly using k means for numerical and squeezer algorithm for categorical, combined the resultant clusters as categorical and finally clustered with squeezer algorithm making the system robust, scalable and efficient.

In future work, we intend to classify each clusters using decision tree induction algorithm to further provide an interpretation to the generated clusters.

REFERENCES

[1] Abraham Silberschatz, Henry F. Korth and Suders S. Han,*Database system Concepts*, fifth edition, McGraw Hill international, 2006.

[2] AsadiSrinivasulu, Ch.D.V.SubbaRao, C. Kishore and Shreyash Raju, *Clustering the Mixed Numerical and Categorical Datasets using Similarity Weight and Filter Method*, International Journal of Computer Science, Information Technology and Management vol.1 No.1-2,2012.

[3] M. A. Honorine, M. Sowjanya and O. Mrudula,*Cluster Ensemble Approach for Clustering Mixed Data*. International Journal of Computer Techniques, vol.2,No.5, p.43-50,2015.

[4] ZhexueHuang,*Clustering Large Datasets with Mixed Numerical and Categorical Values*, 1997.

[5] G. Karypis, E. H. Han, and V. Kumar, *CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling*, http://www.lsi.upc.edu/~bejar/amlt/material_art/DMclustering karypis99chameleon.pdf.

[6] J.MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*. In Proceedings for the 5th Berkeley Symposium on Mathematical Statistics and Probability, p. 281-297, 1967.

[7] Milton Scott, Heinz W. Schmidt,*Dynamic Dispatch in Object-Oriented Languages* (Technical report). TR-CS-94-02. Australian National University. *CiteSeerX*: *10.1.1.33.4292*. *1994*.

[8] Ming-Yi Shih, Jar_Wen Jheng and Lien-Fu Lai, *A Two-Step method for Clustering Mixed Categorical and Numeric Data*, Tamkang Journal of Science and Engineering, vol.13,No.1,2010.

[9] Rafsanjani Kuchaki M., Varzaneh Asghari Z., Chukanlo Emami N., *A Survey of hierarchical Clustering Algorithms*, The Journal of Mathematics and Computer Science vol. 5, No. 3. 2012.

[10] Reddy M. V. Jagannatha and Kavitha B., *Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and filtered Method*, International Journal of Database Theory and Application vol.5, No.1, 2012.

[11] Oded Maimon and Lior Rokach,*Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing Co. Pte Ltd, 2007.

[12] Teknomo Kardi, *K-means Clustering Tutorials*. http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm

[13] Prajapati Madhavi and Dhobi,*Clustering Method for Mixed Categorical and Numerical data*, IJARIE vol. 2 No.3.2016.

[14]S. K. Singh, *Database Systems: Concept, Design and Applications*, Dorling Kindersley (India) Pvt. Ltd., Pearson.2006.

[15] Shi-Hua Liu, Liang-Zhong Shen and De-cai Huang, *A Three-stage framework for clustering mixed data*, WSEAS TRANSACTION on SYSTEMS E-ISSN 2224-2678, vol. 15, 2016.

[16] Sugana J. and Selvi Arul M,*Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data*, International Journal of Computer Application vol.42, No.3, 2012.

[17] Zengyou He, Xu X and S. Deng,*Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach*, Arxiv preprint cs/0509011,2005.

[18] ZengyouHe, Xu Xiaofei and Deng Shengchun,*Squeezer, An Efficient Algorithm for Clustering Categorical data*, J. Comput. Sci. & Technol. Vol.17, No.15,p 1-14, 2002.

[19] A. Topchy, A. Jain, and W. Punch. *A mixture modelfor clustering ensembles*, inSDM, 2004.

[20] S.Sarumathi, N.Shanthi, G.Santhiya, *A Survey of Cluster Ensemble*,International Journal of Computer Applications,Vol.65, No.9, 2013.

[21] Nisha Rani, and Yamini Chouhan. *Combining and Analyzing Apriori and K-Means Algorithms for Efficient Data Mining on the Web*,International Journal of Computer Trends and Technology (IJCTT) Vol.23, No.1, 31-34, published by Seventh Sense Research Group, 2015.

[22] M. Karthikeyan, *Semi Supervised Document Classification Model Using Artificial Neural Networks*, International Journal of Computer Trends and Technology (IJCTT) vol.34, No.1, 52-58, published by Seventh Sense Research Group, 2016.

[23] K. Kavitha, *Pertaining the Concept of Risk Evaluation and Prediction for Multi-Dimensional Clustering*, International Journal of Computer Trends and Technology (IJCTT) Vol.32 No.1, 14-16, published by Seventh Sense Research Group, 2016.