Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

EXTENSION OF K-MEANS ALGORITHM FOR CLUSTERING MIXED DATA

[1]F. E. Onuodu, E. O. Nwachukwu [2], O. Owolabi [3]

[1, 2] Department of Computer Science,
University of Port Harcourt,
P.M.B 5323, Choba, Port Harcourt

[3] Department of Computer Science
University of Abuja, Nigeria.

Email: friday.onuodu@uniport.edu.ng[1], enoch.nwachukwu@uniport.edu.ng[2], doowo@yahoo.com[3]

Received: 21-06-14
Accepted: 20-08-14

ABSTRACT

In this work, a new hybrid method has been proposed which extends K-means algorithm to categorical domain and mixed-type attributes. Also proposed is a new dissimilarity measure that uses relative cumulative frequency-based method in clustering objects with mixed values. The dissimilarity model developed could serve as a predictive tool for identifying attributes of objects in mixed datasets. It has been implemented using JAVA programming language and MATLAB. Experiments on real-world datasets show that the new hybrid algorithm is more efficient and more robust when compared with existing ones in terms of accuracy and time complexity. This tool can be used in a variety of applications such as in agro-based industries, in clinical datasets and in general information retrieval system (IRS). The new method has been applied on agro-based datasets of soybean and yeast for forming clusters that could help farmers in the management of crop pests.

Key words: Mixk-meansXFon, Clustering, Mixed data.

INTRODUCTION

There is an increasing interest in the development of data mining applications using cluster analysis. The major problem with mixed data is clustering massive datasets. However, these algorithms appear to have concentrated on numeric data only and cannot work for large objects with mixed datasets. The data mining community has put in a lot of efforts in developing fast algorithms for clustering very large datasets. Some of these algorithms were modifications and extension of the existing clustering methods. Searching for useful information from a large collection of data, popularly called data mining, has become very important in the information age. An interesting area of research that is concerned with data organization in terms of meaning, structure and interpretation is data clustering. Data clustering is the process of grouping objects with high degree of relationships together with objects that have low degree of relationships. This means that objects found in a group are highly similar and share common attributes that are distinct from objects in other groups. Clustering itself is aimed at finding useful groups of clusters where usefulness is defined by the aims of the data analysis. In a way, a good cluster is determined by how well groups of a particular group of objects

are separated from other groups such that clusters or groups cannot overlap one another.

The problem of data clustering is to know accurately the number of clusters that can be formed out of a given set of data and then choosing the right clustering models for it. This is because different data from different domains can have similar internal structures; and this could lead to wrong estimates of clusters and the grouping of wrong objects to groups that they do not belong. The issue of high dimensionality in a huge dataset can as well be a very big problem; due to the cost of operations involved when forming clusters. There are several clustering algorithms that have been developed to solve these problems. But as some of these algorithms tend to solve these problems, they also create new problems to be solved. K-means is one such algorithm that is very popular in the area of clustering because it is easy to implement and at a low computational cost.

In this work, a new hybrid method that extends K-means algorithm to categorical domains and mixed type attributes has been developed. The new method, MixK-meansXFon, matches different datasets with different clustering algorithms. Also proposed is a new dissimilarity measure that uses relative cumulative frequency-based method in clustering objects with mixed values. The dissimilarity model developed could serve as a predictive tool for identifying attributes of objects in mixed datasets. The proposal was implemented using JAVA programming language and MATLAB for data extraction and graphical multi-domain simulation respectively. Structured System Analysis and Design Methodology (SSADM) were used in this approach and two datasets obtained from UCI repository were used to demonstrate the clustering performance of the algorithm. Experiments on real-world datasets show that the new hybrid algorithm is more efficient and more robust when compared with existing ones in terms of accuracy and time complexity. The quality of clusters formed using the new method was also quantified by the sum of squared errors (SSEs) value of 0.15 when K=8 with O (N^2) running time. The use of Graphical User Interface in this work gives an overall visualization of how the clusters are formed. This tool can be used in a variety of applications such as in agro-based industries, in clinical datasets and in general information retrieval system (IRS). The new method has been applied on agro-based datasets of soybean and yeast for forming clusters that could help farmers in the management of crop pests.

Asadi et al. (2012) presented a cluster of mixed numeric and categorical datasets in an efficient manner. They used a clustering algorithm based on similarity weight and filter method paradigm, that works well for data with mixed numeric and categorical features. Although the approach was very efficient in solving any number of dimensions, the problem is that they could not match the different clustering datasets with different algorithms. Ahirwar, R. (2014) also presented an efficient algorithm that made use of Divide and Conquer techniques to cluster large datasets. He compared the performance of the proposed algorithm with the F-measure, purity and Entropy and the results show that the approach used was much better than the existing algorithms. Although the approach was very efficient in identifying the data points and assigning of the data points to the best clusters, the problem is that he could not read the entire data files at once. Pham, D. et al. (2011) presented a new algorithm

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

to cluster datasets with mixed numerical and categorical values. The new algorithm combined the advantages of most recently introduced population-based optimization algorithm called Bees algorithm (BA) and K-prototypes. Although RANKPRO algorithm was more efficient than K-prototypes algorithm for a specific dataset, yet the K-prototypes algorithm converged to a local minimum very fast in a few iterations. San et al. (2004) presented the notion of “cluster centers” on datasets of categorical objects and showed how this notion could be used for clustering problems of categorical objects as a partitioning problem. The experiments show that the proposed algorithm gave better results and appear to be more stable than k-modes algorithm but the problem is that they could not combine the proposed algorithm with the k-means algorithm in a similar manner as was done in (Huang, 1997; Huang, 1998), when k-means paradigm was applied to cluster mixed datasets.

MATERIALS AND METHODS

The clustering architecture is shown in figure 1, which actually suggest (or explain) the proposed hybrid architecture and model for clustering mixed data.

Analysis

The proposed hybrid algorithm, MixK-meansXFon (Clustering Mixed-data as Extension to K-means)**,** is a more preferred way of integrating the k-means and the extended k-modes algorithms into the MixK-meansXFon algorithm used for clustering the mixed data. The model will be more useful because most objects that occur frequently in real world databases are mixed-type values. The dissimilarity measure between two mixed-type objects X and Y are described by their attributes Ar_1, A^r_2,.., A^r_p and $A^c_{p+1,\dots}$ A^c_m. This can be designated as:

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} rcf\delta(x_j, y_j)$$

…(1)

where first term is numeric attributes and the second term is the simple matching dissimilarity measure on the categorical attributes. The weight γ is used here to avoid favoring the categorical or numerical attribute. The influence of γ in the clustering process is discussed in (Huang, 1997)

The general algorithm for given set D = $\{X_1, . . . ,Xn\}$ of n numerical data objects, a natural number $k \leq n$, and a distance measure d, the k-means algorithm is aimed at finding a partition C of D into k non-empty disjoint clusters $C_1, . ., C_k$ with $C_i \cap C_j = \phi$; and $\cup^{k}{}_{i=1} C_i = D$ such that the overall sum of the squared distances between data objects and their cluster centers are minimized. Mathematically, if indicator variables w_i, is used, which takes value 1 if object X_i is in cluster C_l, and 0 otherwise, then the problem can be stated in terms of a constrained non-linear optimization problem:

Minimize

$$p(W,Q) = \sum_{i=1}^{k}\sum_{i=1}^{n} w_{i,l}, d(X_i, Q_l),$$

...............…………………………………...(2)

Subject to

$$\sum_{l=1}^{k} w_{i,l} = 1, 1 \leq i \leq n,$$

………...........……………………(3)

$w_{i,l}$ ε {0, 1}, $1 \leq i \leq n$, $1 \leq l \leq k$, where Q = $[w_{i,l}]_{n\times k}$ is a partition matrix, Q = $\{Q_1, . . . ,Q_k\}$ is the set of cluster center, and d (.,.) is

the squared Euclidean distance between two objects.

With the modifications made, the problem of clustering categorical data as a partitioning problem in a fashion similar to k-means clustering can be formulated. Assume that a data set, $D = \{X_1, \ldots, X_n\}$ as categorical objects to be clustered, where each object $Xi = (x_{i,1}, \ldots, x_{i,m})$, $1 \leq i \leq n$ is described by m categorical attributes.

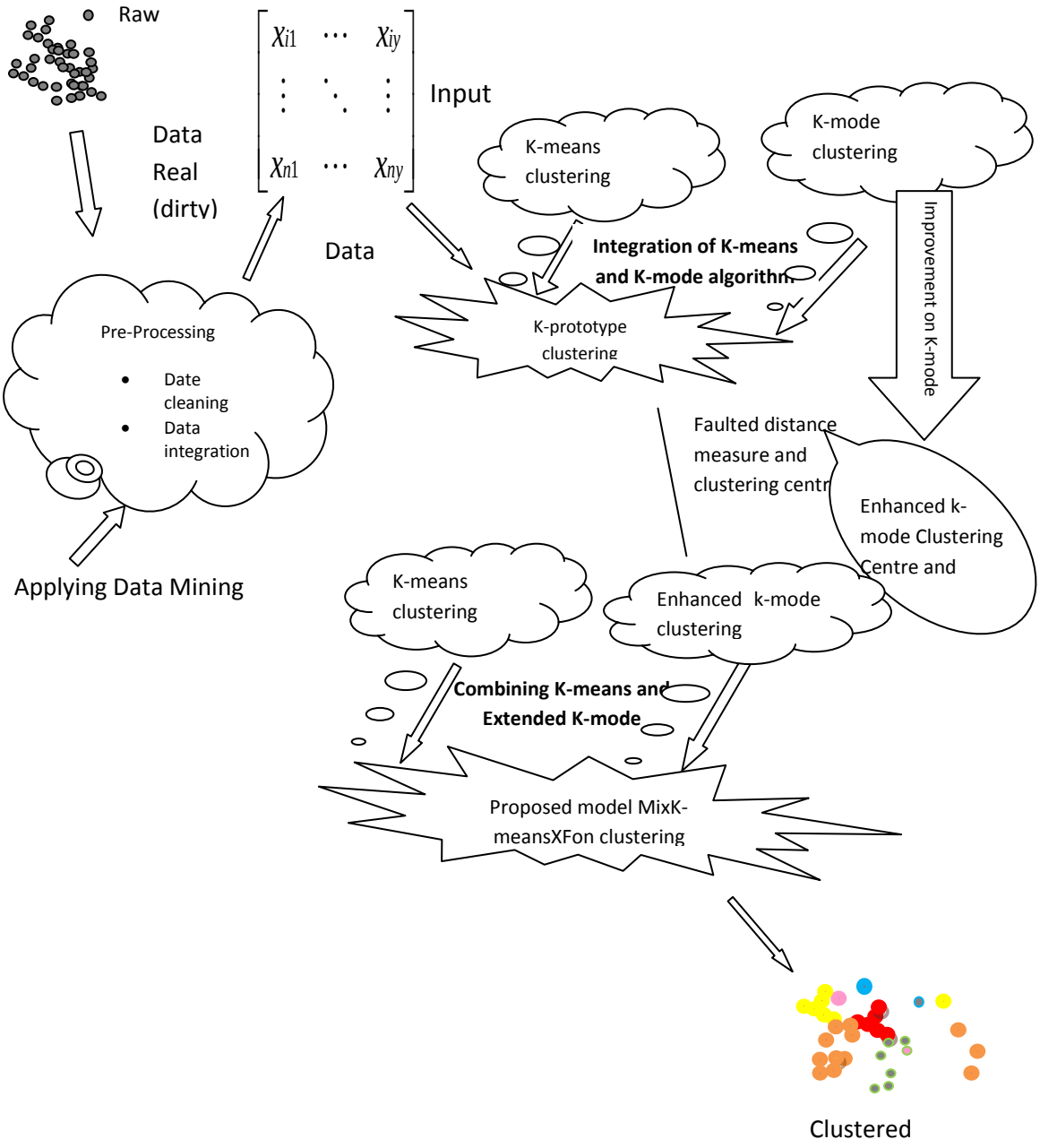

Figure 1: Proposed Hybrid Architecture

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
 ISSN 1118 - 1931

Then, the problem can be mathematically stated as follows:
Minimize

$$p(W,Q) = \sum_{i=1}^{k}\sum_{i=1}^{n} w_{i,l}, d(X_i, Q_l),$$

……(4)

Subject to

$$\sum_{l=1}^{k} w_{i,l} = 1, 1 \le i \le n,$$

…… (5)

$w_{i,l} \varepsilon \{0, 1\}$, $1 \le i \le n$, $1 \le l \le k$,
where $W = [w_{i,l}]_{n\times k}$ is a partition matrix, $Q = \{Q_1, \ldots, Q_k\}$ is the set of representatives, and $d(X_i,Q_l)$ is the dissimilarity between object X_i and representative Q_l. In the same way as in the k-modes algorithm proposed in (Huang, 1998), the algorithm for clustering categorical data and mixed-type objects can be introduced by modifying the cost function of equation (1) as follows:

$$P(W,Q) = \sum_{l=1}^{k}\left(\sum_{l}^{n} w_{i,l}\sum_{j=1}^{p}(x_j - y_j)^2 + \gamma\sum_{I=1}^{n} w_{i,I}\sum_{j=p+1}^{m} rcf\,\delta(x_j, y_j)\right)$$

……(6)

Let

$$P_l^r = \sum_{j=1}^{n} w_i\, j \sum_{j=1}^{p}\left(x_{i,j} - q_{i,j}\right)^2$$

..... (7)

And

$$P_1^c = \gamma\sum_{i-1}^{n} w_{i,l}\sum_{j=p+1}^{m} rcf\,\delta\left(x_{i,j} q_{i,j}\right)^2$$

..... (8)
Equation (4) can then be rewritten as:

$$P(W,Q) = \sum_{l=1}^{k}\left(P_l^r + P_l^c\right)$$

..... (9)

Since both P^r_l and P^c_l are non-negative, minimizing P(W; Q) is equivalent to minimizing P^r_l and P^c_l for $1 \le i \le k$ and rcf is the relative cumulative-frequency. This method was simulated in MATLAB 7.7 (R2008b) version and JAVA for data extraction. The algorithms used are K-means, Extended K-modes as well as the Hybrid as explained in the next section.

K-means Algorithm
The basic k-means clustering technique is described as follows:
Step 1: Select k points as the initial centroid.
Step 2: Assign all points to the closest centroid.
Step 3: Re-compute the centroid of each cluster.
Step 4: Repeat steps 2 and 3 until the centroid can no longer be changed.
The number of cluster K will be determined first and the centroid or center of these clusters can then be assumed. Take any random object as the initial centroid or the first k object can also serve as the initial centroid. Then, the k-means algorithm will do the three steps below until convergence, iterate until stable (i.e. when no changes in the each group):

Extended K-modes Algorithm
The Extended K-modes algorithms (Aranganayagi, et al. 2010) are as follows:

1. Initialize Modes of K clusters
2. Compute the dissimilarity between the object and the modes of the clusters. Place the object in the cluster which results in minimum dissimilarity. Update the mode of the cluster.
3. After all objects have been allocated to the respective cluster, retest the object with new modes and update the clusters
4. Repeat steps (2) and (3) until there is no change in clusters.

The k-modes algorithm has the following modifications to the k-means algorithm. These include:

(i) Using a simple matching dissimilarity measure for categorical objects,
(ii) Replacing the means of clusters with the modes, and
(iii) Using a frequency-based method to find the modes.

These modifications have removed the numeric-only limitation of the k-means algorithm but maintain its efficiency in clustering large categorical data sets (Huang, 1998).

New Hybrid Pseudocode (Modified Part)

```
// store the datasets in a matrix object using an array data structure
Matrix = dataset [ ][ ]
// iterate through the objects in the array using the Nested For Loop construct
For i = 1 to NumberOfObjects
// compute the distance measure for the dataset stored as a matrix and add the weight, λ
distanceMatrix = distanceMeasure(matrix) + λ
// loop through the cluster number and call the dissimilarity function to compute dissimilarity measure for the mixed dataset (categorical and numerical)
For j = 1 to ClusterNumber
dissimilarityMatrix = dissimilarityMeasure(numerical, categorical)
// call the cluster procedure to return the cluster of the relative cumulative frequency (rcf) of the dissimilarity matrix
cluster(rcf(dissimilarityMatrix))
// compute the relative cumulative frequency of the dissimilarity matrix of the mixed dataset
Procedure rcf(dissimilarityMatrix)
// assign the frequency of the dissimilarityMatrix to a variable clusterfrequency

clusterfrequency = freq(dissimilarityMatrix)

// compute the rcf using the cluster frequency and frequency of the dissimilarity matrix

rcfdissimilarityMatrix = clusterfrequency / freq(dissimilarityMatrix)
return rcfdissimilarityMatrix

end

// procedure to compute the clusters using the rcf of the dissimilarity matrix as a parameter
procedure cluster( rcfdissimilarityMatrix)

// initialize cluster count to zero
clusterCount = 0

// loop through the items in the array

for i = 1 to numberOfCategoricalAtrrib

for j = 1 to numericalAttrib

// Assign the attribute to the cluster
cluster[i][j] = rcfdisMatrix

// increment cluster count
clusterCount = clusterCount + 1

// terminate the loop

end for

end for

return

end
```

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Design

The design of a generic hybrid method is shown in figure 2, which simply explains the processes involved in clustering mixed data. The datasets are of two parts; the numerical and categorical attributes. It starts from cleaning the data to remove noise and any inconsistencies. It then goes on to data integration and data selection by retrieving relevant data for analysis. The process continues with data transformation that forms a suitable mining for carrying out summary or aggregate operations. The next stage determines the distance measure for attributes of both numerical and categorical datasets. The computed similarity or dissimilarity of attributes for numerical and categorical are then combined into single dissimilarity for mixed data. The combined clustering using K-means and Extended K-modes for the dissimilarity of mixed attributes is then performed to produce the final clustered results. The k-means and k-modes methods can only cluster numerical values and categorical attributes respectively. The new hybrid method makes use of these two algorithms after a few modifications and extended the k-means paradigm to categorical domain and mixed-type attributes. The hybrid method, MixK-meansXFon, matches different clustering datasets with different algorithms, which appear to be lacking in Asadi (2012). From figure 1, the existing distance measure has been improved by making use of Minskwoski distance, which is a generalization of Euclidean and Manhattan distances. The notion of relative cumulative frequency (rcf) method was also adopted in this work because the frequency-based method does not give an even distribution of how clusters are formed. This rcf plays a significant role to the clustering in the sense that the data points are evenly spread out across the plot and this reduces the risk of producing bad clusters.

RESULTS

The clustering algorithm was demonstrated with some standard datasets like soybean and yeast. The new hybrid algorithm was then applied on these datasets to produce clustered results. From figures 3, 4, 5 and 6, the program produced clustered results for k=8, 9, 10 and 12. That is, the sample output appears in the form: cluster 1, cluster 2, ..., cluster 10 and cluster 12. The results were used to test for the accuracy and efficiency of the new algorithm and then compared it with results from k-prototypes developed by Zhexue (1998). The results were used to benchmark the accuracy and efficiency of the enhanced k-means algorithm developed by Abdul et al (2009). The results obtained show that the new algorithm is more efficient and more robust when compared with existing ones. The use of Graphical User Interface (GUI) in this work also gives an overall visual impression of how the clusters are formed, which is quite different from the command line interface used in some previous literatures.

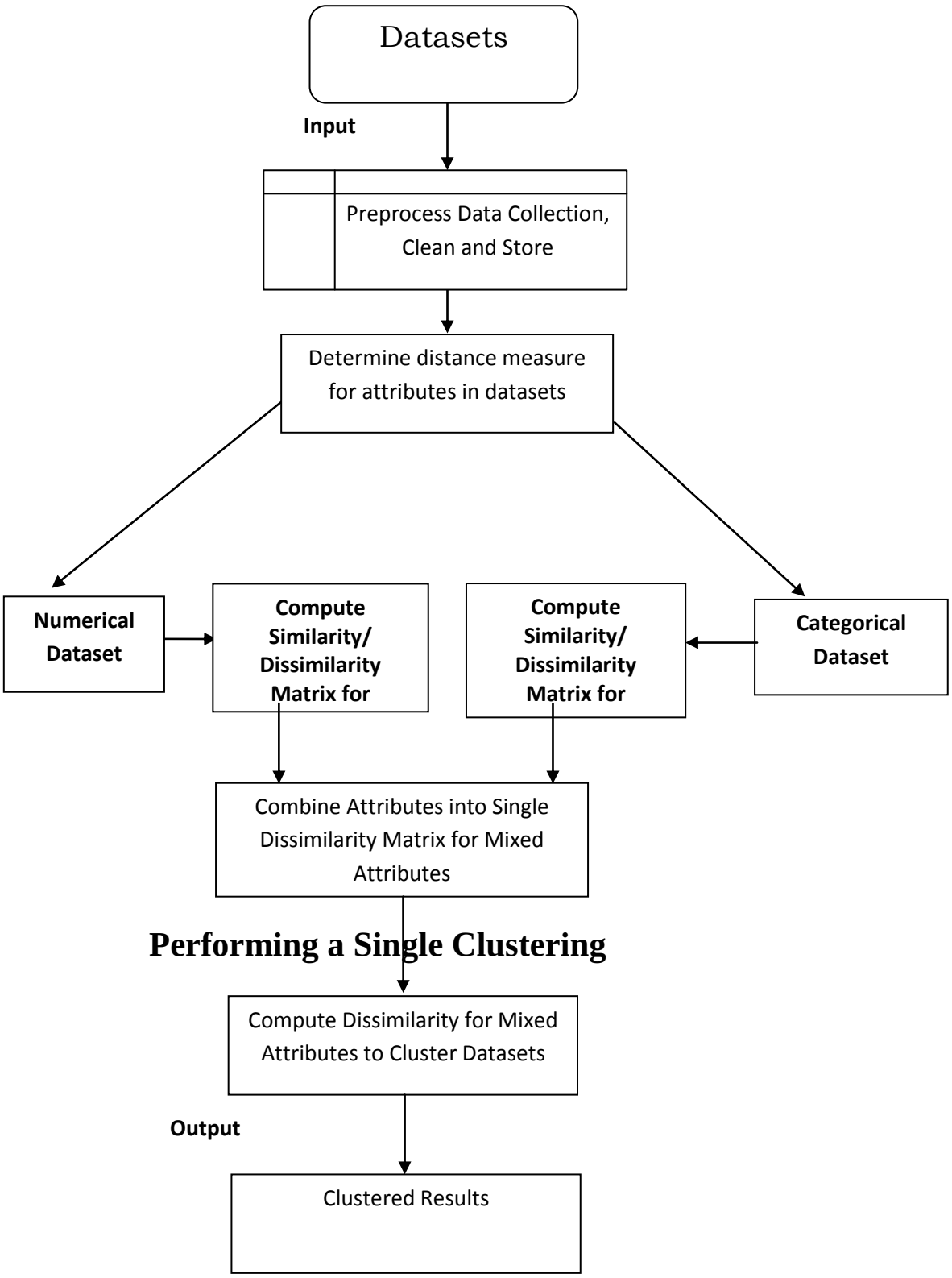

Figure 2: Design of a Generic Hybrid Method for Clustering Mixed Data

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

DISCUSSION

Table 1 is the comparative analysis of the new hybrid method and other methods in terms of accuracy and efficiency. The accuracy measure for K-modes, K-prototypes, Extended K-modes and the new hybrid algorithm were illustrated. The bar chart in Figure 7 shows that the new algorithm has accuracy measures of 0.86 soybeans and 0.84 of yeast, which shows that it is more efficient and more accurate than the traditional K-prototypes 0.69 soybean and 0.74 of yeast, Extended K-modes has 0.83 and K-modes of 0.37 for soybeans only. A few researches on yeast may have accounted for no results in the case of K-modes and extended K-modes. It is also observed that when the same set of modes is applied for these algorithms, the new algorithm outperforms the original K-prototypes and other algorithms as shown in figure 7 and 8 respectively.

Performance of New hybrid Algorithm on Different Mixed Datasets and Sum of Squared Errors (SSEs)

The performance evaluation of the new hybrid method has been applied on different mixed datasets and the SSEs value has been shown in table 2 and compared with Ming-Yi et al., (2010). The datasets were obtained the UCI Machine learning Repository Irvine Asuncion et al. (2013).

Table 1: Comparative Analysis of Results of Hybrid Algorithm and K-prototypes

S/N	Datasets	New Hybrid Algorithm (MixK-meansXFon, 2014)	K-Prototypes (Zhexue, 1998)	Extended K-modes (Aranganayagi, 2010)	K-modes
1	Soybean	0.86	0.69	0.83	0.37
2	Yeast	0.84	0.74	-	-

SNAPSHOT OF SOYBEAN CLUSTERING

Figure 3: **Clustered Output from Soybean dataset**

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
 ISSN 1118 - 1931

SNAPSHOT OF MATLAB SIMULATION OF SOYBEANS

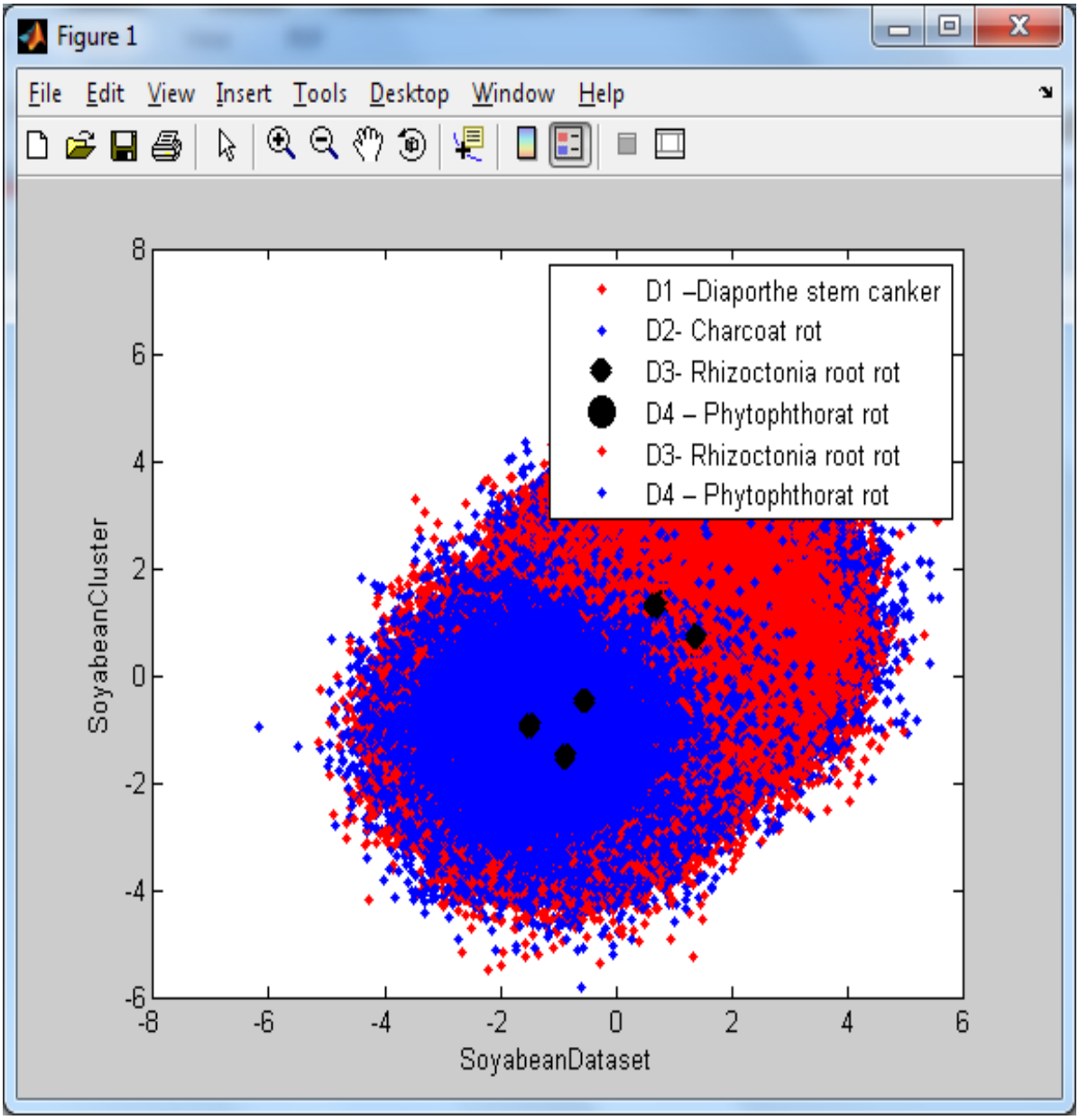

Figure 4: MATLAB output of soybean clusters.

SNAPSHOT OF YEAST CLUSTERING

Figure 5: Clustered datasets of yeast

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
 ISSN 1118 - 1931

SNAPSHOT OF MATLAB SIMULATION OF YEAST

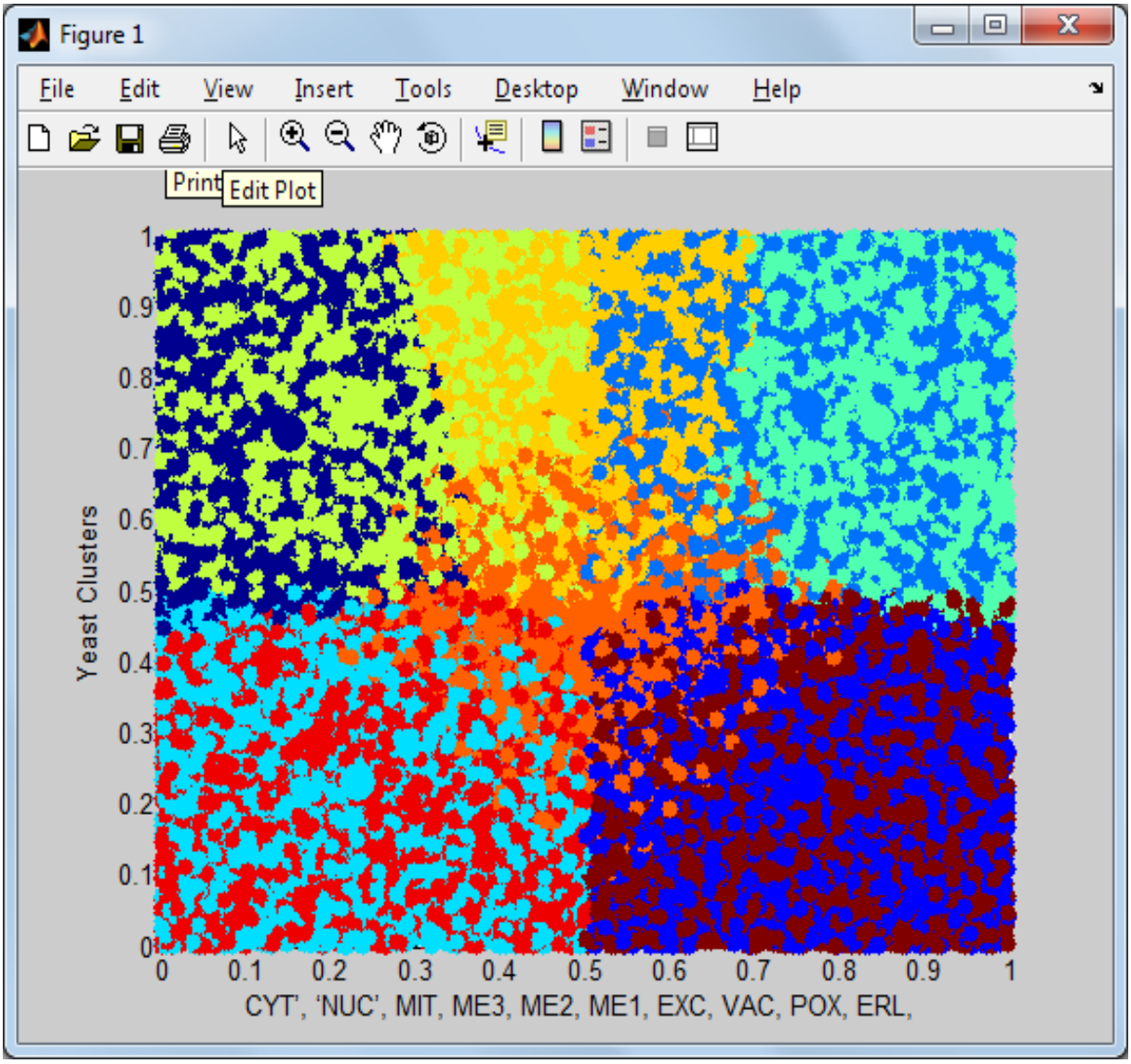

Figure 6: MATLAB output of yeast clusters.

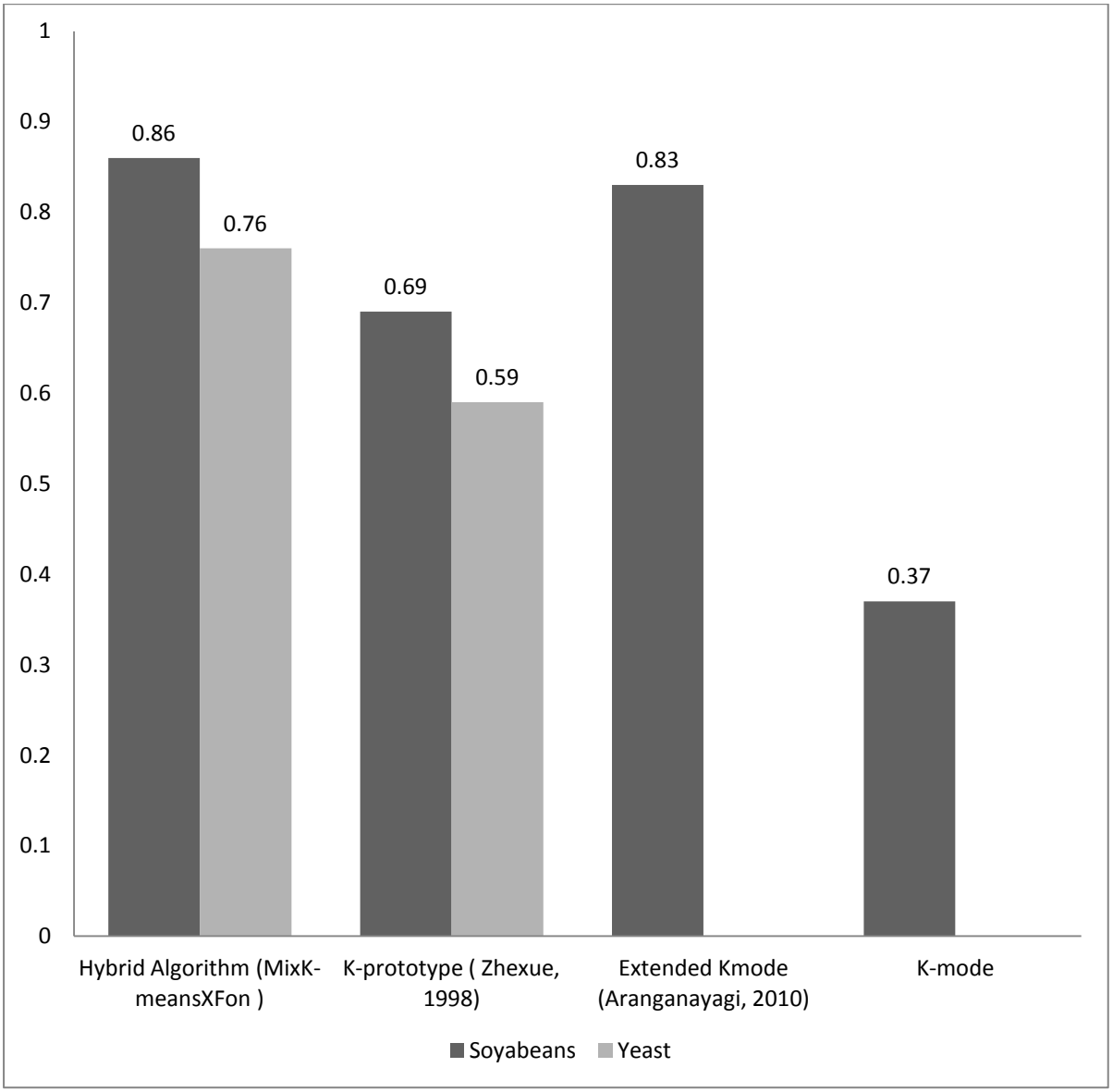

Figure 7: Graphical representation of the **comparative results of k-modes, k-prototypes, extended k-modes and hybrid algorithm**

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

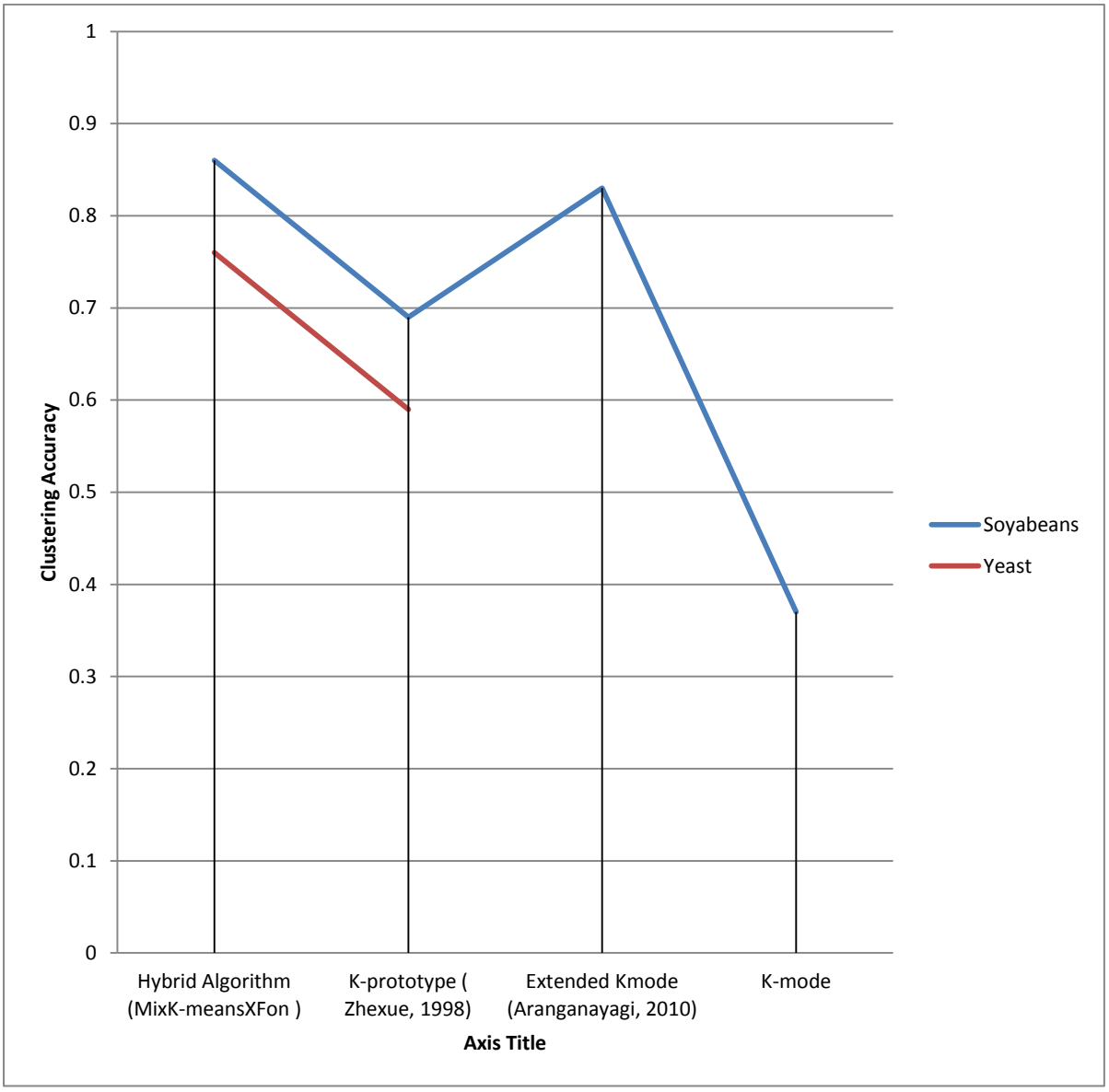

Figure 8: Shows that the Hybrid Algorithm is superior to the traditional K-prototypes clustering algorithm.

SNAPSHOT OF CLUSTERING OF MIXED DATASETS OF HEART DISEASE, CREDIT APPROVAL AND SOYBEAN AND YEAST.

Figure 9: Shows the clustering of mixed datasets of Heart disease, Credit approval and Soybean and Yeast datasets using new hybrid algorithm

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

Table 2: SSEs Values of MixK-meansXFon on Different Mixed Datasets

Cluster Number (K)	Heart Disease (Ming-Yi et al., 2010)	Credit Approval (Ming-Yi et al., 2010)	Soybean and Yeast (New Hybrid Algorithm)
2	0.33	3.63	1.75
4	0.13	0.86	0.54
8	0.07	0.18	0.15

The graph shows a plot of the sum of squared errors (SSEs) against the cluster of different datasets (Heart disease, credit approval and mixed soybean and yeast datasets). The results from the graph in figure 9 show the performance of the MixK-meansXFon algorithm on different datasets. It can be seen that the new algorithm performs well on the mixed datasets albeit with some differences. The Heart disease dataset has the best performance of the three with very low sum of squared error (SSEs) while the mixed soybean and yeast dataset performed averagely with SSEs value of 0.15 when K=8. The credit approval dataset performed strongly with higher values of the cluster number (K) but for very low values, it performed poorly. In the overall performance, the new hybrid algorithm is a more promising method and more robust when applied on different mixed datasets. The time complexity is N^2, ignoring the constant term and focusing on the dominant term in the expression. The order of growth for the time complexity is quadratic for the input size. That is, $O(n) = N^2$, where N is the total number of objects. This indicates that the new hybrid algorithm has $O(N^2)$ running time, which is good and justifiable considering the massive datasets involved.

CONCLUSION

In this work, the hybrid algorithm, MixK-meansXFon, was proposed for clustering mixed data. JAVA and MATLAB were used for implementation. The clustered accuracy of the new algorithm was compared with existing algorithms like K-modes, K-prototypes (1998) and extended K-modes (2010). The new algorithm is more efficient and more robust than K-modes, K-prototypes and extended K-modes. The limitation of the new algorithm is that the value of *k*, which is the number of desired clusters, is still required to be given as input, regardless of the distribution of the data points in clustering mixed data.

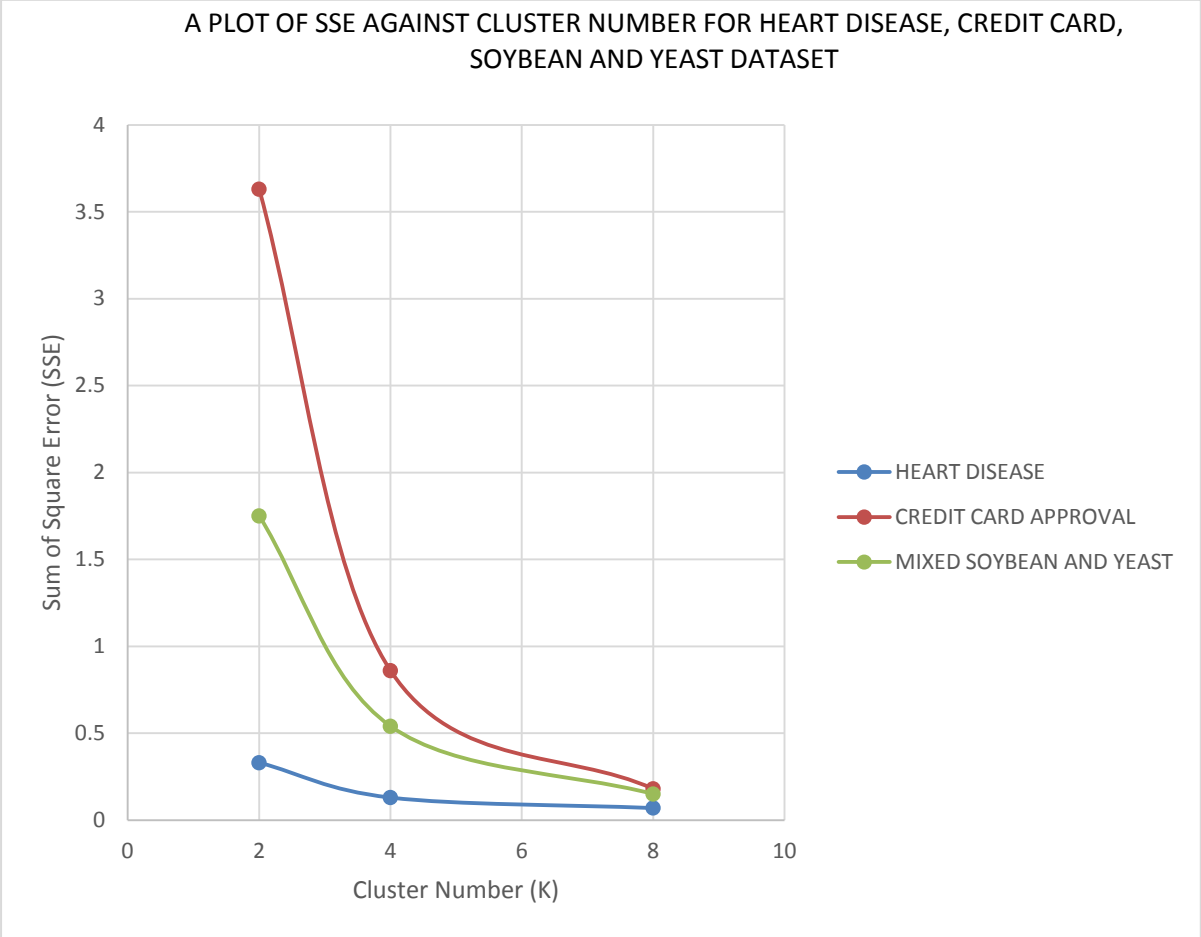

Figure 10: Graph of SSEs against Cluster Number for Heart Disease, Credit Approval, and Soybean and Yeast datasets

REFERENCES

Asadi, S., Subba, R. C.D.V., Kisshore, C. and Raju, S. (2012) Clustering the Mixed Numerical and Categorical Datasets Using Similarity Weight and Filter Method. International journal of Computer Science Information Technology and Management, 1:1-2

Abdul, K.A and Sebastian, M.P. (2009), Improving the Accuracy and Efficiency of the K-means

Clustering Algorithm, Proceedings of the World Congress on Engineering, London, U.K 1:1-3,

Ahirwar, G. (2014), A Novel K-means Clustering Algorithm for Large Datasets Based on Divide and Conquer Techniques, Pradnyesh.J.Bhisikar (IJCSIT) International Journal of Computer Science and Information Technologies, 5(1): 301-305.

Arangnayagi, S. and Thangavel, K. (2010) Extended K-modes with Probability Measure International Journal of Computer Theory and Engineering, 2(3): 431-435

Scientia Africana, Vol. 13 (No.2), December 2014. Pp180-198
 ISSN 1118 - 1931

Asuncion, A. and Newman, D.J.(2013) UCI Machine learning Repository Irvine, CA: University of California, School of Information and Computer Science.

Huang, Z.(1997) Clustering Large datasets with mixed numerical and Categorical values, Data mining Knowledge Discovery, Techniques and Applications (H. Lo H.Lon, Eds) Singapore World Scientific, 21-34.

Huang, Z.(1998) Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values, Data mining Knowledge Base. *Discovery* at Netherlands 2(2): 283-304.

Jain, A.K, and Dubes, R. (2000). Clustering Techniques: The User Dilemma.

Pham, D-T., Suarez-Alvarez and Prostov, Y.I.(2011), Random Search With K-prototypes Algorithm for Clustering Mixed Datasets, Journal of the Royal Society, proceeding R.Soc. A doi:10.1098/rspa.2010.0594.

San, O.M, Huynh, V.N. and Nakamori, Y. (2004), An Alternative Extension of the K-means

Algorithm for Clustering Categorical Data, Japan Advanced Institute of Science and Technology, Int.Journal. Math.Comput. Sci, 14(2): 241-247.