Scientia Africana, Vol. 13 (No.2), December 2014. Pp225-241
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 - 1931*

A HYBRID APPROACH ON THE MANAGEMENT OF CROP PESTS

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

Department of Computer Science,
University of Port Harcourt, P.M.B 5323, Choba,
Port Harcourt, Rivers State, Nigeria (1, 2)

Department of Computer Science,
University of Abuja, Nigeria.(3)
***Email**: friday.onuodu@uniport.edu.ng[1], enoch.nwachukwu@uniport.edu.ng[2], doowo@yahoo.com[3]*

Receieved: 20-06-14
Accepted: 28-08-14

ABSTRACT

In this work, a new hybrid clustering approach is presented that could be applied in mitigating against the spread of pests by first identifying the region in the field in which crops are affected and then performing a clustering algorithm to cluster and separate affected crops. Also discussed is a new dissimilarity model that could serve as a predictive tool for identifying attributes of objects in mixed datasets. This algorithm has been implemented using JAVA and MATLAB. 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. The model developed could be beneficial to Nigerian farmers and the Agro-based industries, particularly those involved in crop scouting.

Key words: Data Analysis, clustering, dissimilarity model, crop pests.

INTRODUCTION

The data miners are searching for fast algorithms that will be efficient in clustering complex large datasets. However, the traditional methods of data analysis in agriculture appear to be purely dependent on statistics. Data mining is a promising modern data analysis technique. Agriculture is the largest sector of the Nigerian economy with Gross Domestic product (GDP) contribution of about 40%. Furthermore, Nigeria has over 80 million hectares of arable land, which accounts for about 23% of arable land across West Africa. In agro-based industries, the volume of data is enormous as such when converted to information, will be very useful for many purposes (Babajide et al, 2012).

Most agricultural crops have been threatened by a wide range of pests. Today, these pests can either damage crops or wipe out an entire harvest. It is estimated that up to 42% of the world’s total agricultural crop is destroyed by pests annually (SPM, 2000). The traditional methods of identifying pathogen have always been by general symptoms, by isolating and culturing of the organisms, and by morphological and biochemical tests, which happen to be the main goal of the integrated pest management. Integrated pest management (IPM) is primarily concerned with crop production using effective, stable, and long-lasting crop protection components that minimizes the negative side effects of pest control actions. IPM recognizes that farmers' knowledge is the key to success and not the technology only. In a more traditional sense, IPM has been thought of as the use of multiple tactics to optimize control, but slowly that vision has changed to accommodate the integration of all pest management tactics for crop pests. The triangle of IPM includes the host, pathogen and the environment. The three factors are necessary for the development of plant diseases, thus, disease can also be affected by altering any of these three factors (Kogan, 1998).

There is a growing need for agricultural researchers to use proper procedures and knowledge of economic thresholds to ensure that farmers use pesticides properly and realize maximum profits for their investment. The problem of classification in a typical crop disease datasets is a very big task. The form they are categorized will not adequately suggest the best treatment for the diseases. It then follows that, what could work for one crop pest may act poorly for others. The soybeans pest management in North Carolina has reported that pest can damage the crop or interfere with its growth, thus causing a huge loss (SPM, 2000). The clustering of agro-based datasets as suggested in this work will serve as means to mitigate the spread of pest by first identifying the region in the field in which the crops are affected and then performing the clustering algorithm to cluster and separate affected crops. The work will be of interest to agro-based organizations involved in managing of crop pests and probably suggest the best way to save cost. A few literatures have been reviewed.

Ramesh, et al. (2013) implemented a parallel k-means algorithm to cluster large datasets of agricultural databases. They experimented with agricultural datasets due to limited researches that have been done in agricultural field. Their aim was to improve the time complexity inherent in k-means algorithm when applied on large datasets since it used to be computationally expensive. Although, the performance of the algorithm proved to be better in terms of efficiency and time complexity compared with the normal sequential k-means, the problem is that the processing speed was disturbed with the bandwidth of the network available. Dubey et al. (2013) presented a novel defect segmentation of fruits based on color features with k-means clustering unsupervised algorithm. They used color images of apple fruits and this was done in two stages. At first, the pixels were clustered based on color and spatial features. Secondly, the clustered blocks were merged to a specific number of regions. These two procedures were able to increase the computational efficiency, at the same time, avoiding feature extraction for every pixel in the image of apple fruits. Although their results show that the proposed k-means-based defect segmentation was able to accurately segment the defected area of fruits present in the image, the problem is that they could not automatically determine the number of clusters required to segment the defects more accurately. Kolhe, et al. (2010) implemented a new method based on a combination of Supervised and Unsupervised learning of clustering data without any preliminary assumption on the cluster shape. This was achieved by extracting the dissimilarity relations directly from the available data. A feed forward neural network (FFNNs) was trained with a few known dissimilarity degrees between pattern pair. The neural dissimilarity relation served as input to the fuzzy clustering algorithm. The results show that the clusters were formed correctly when the attributes are petal length verses petal width having accuracy of 92.38% for 35% of patterns. Although their results were encouraging and trustable, they could not separate Iris virginica from Iris versicolor when using a typical Iris flowering plant as case study. Kleinberg, (2002) came up with the idea of no best clustering algorithm as partially captured by the impossibility theorem, which states that no single clustering algorithm simultaneously satisfies a set of basic axioms of data clustering. He opined that algorithms developed may give best results with one type of data set but may fail or give poor results with data set of other types. Although, there have been many attempts to standardize clustering algorithms in order to obtain algorithms that can perform well under different scenarios, but the fact remains that each algorithm has its own merits and demerits and cannot work for all real situations. The fields where clustering algorithms are used are quite diverse. It becomes difficult to exhaustively list the numerous science fields and applications that have utilized clustering techniques as well as the thousands of published algorithms. Therefore, there is need to extend the new hybrid algorithm to the management of crop pests in a Nigerian situation.

METHOD

Analysis

The proposed hybrid algorithm, MixK-meansXFon (Clustering Mixed-data as Extension to K-means)**,** is a more preferred way of combining the k-means and Extended k-modes algorithms into the MixK-meansXFon algorithm used to cluster the categorical attributes. The proposed system will be more useful since frequently occurred objects in real world databases are mixed-type objects. The dissimilarity between two mixed-type objects X and Y are described by the attributes of Ar$_1$, A^r_2,.., A^r_p and $A^c_{p+1,\dots}$ A^c_m.

This can be measured by:

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

………………………………………… (1)

where the 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 either the categorical or numerical attribute. The influence of γ in the clustering process is discussed in (Huang, 1997). In the same way, the k-modes algorithm proposed by (Huang, 1998) can be introduced for clustering categorical data and mixed-type objects by modifying the cost function of equation (1) as follows:

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

Equation (2) can then be rewritten as:

$$P(W,Q)=\sum_{l=I}^{k}(P_l^r+P_l^c) \text{ -------(3)}$$

the relative cumulative-frequency. The two algorithms namely; K-means and Extended K-modes as well as the new hybrid will be discussed.

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 until convergence, iterate until stable (i.e. when no changes in the each group):

Extended K-modes algorithm

The Extended K-modes algorithm (Aranganayagi, et al. 2010) is 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 the clusters.

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

(iv) Using a simple matching dissimilarity measure for categorical objects,

(v) Replacing the means of clusters with the modes, and

(vi) 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 datastructure
        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
```

Integrating the new hybrid algorithm to the integrated pest management (IPM) program

The integration of a new hybrid, Mixk-meansXFon, clustering approach into IPM program will serve as a better platform (or means) to mitigate the spread of pest by first identifying the region in the field in which the crops are affected and then performing the clustering algorithm to cluster and separate affected crops. To achieve this, two factors were considered namely; crop scouting and economic thresholds. They are explained in the section accordingly.

Crop Scouting

This is part of integrated pest management program used to help farmers identify pest problems and to provoke them for the necessary action if possible. It is therefore advisable that instead of relying on chemicals, the IPM considers all appropriate methods of reducing pest levels to help produce more crops, thereby making profits. The farmers are usually in a haste to recognize what pest can cause, but it is also worth knowing that using pesticide when it is not needed can cause more harm than the profits envisaged for.

Economic Thresholds

The whole idea about economic threshold is to know precisely when to take action against crop pests and how many pests can be present before they can cause economic damage to the crops. Also referred to as ‘Action threshold’ is primarily concerned with the number of pest or level of pest infestation it takes to justify spending money on pesticide treatment. It is also considered a big waste to treat crop pests that may not reduce the quality or yield. It is therefore advisable to allow pest population to build up to a relative large numbers before it can be considered cheaper or economical for the farmer to spend.

Design

The design of the hybrid architecture is shown in figure 1. The datasets are of two parts namely; the numerical and categorical attributes. The k-means and extended k-modes methods can only cluster numerical values and categorical attributes respectively. The proposed hybrid method makes use of these two algorithms. This hybrid method extends the k-means algorithm to categorical domains and mixed type attributes. The new hybrid method, MixK-meansXFon, matches different clustering datasets with different algorithms, which was lacking in a recent work carried out by Asadi (2012). From the generic architecture, the distance measure used was faulted because its distribution was of frequency-based method. A new dissimilarity measure was proposed that uses a relative cumulative frequency-based method (rcf) in clustering objects with mixed datasets.

How similarity/dissimilarity measures work for mixed datasets

Table 1 is a typical mixed dataset comprising both categorical and numerical attributes from numerical yeast dataset and categorical soybean dataset. There are six (6) instances of yeast and six (6) instances of soybean but all the classes and attributes are combined into a single table as shown in table 1.

The Sequence Name represents the class of yeast while the attributes are mcg, gvh, alm, mit, erl
where

i. mcg : Mc Geoch’s method for signal sequence recognition
ii. gvh: Von Heigne’s method for signal sequence recognition
iii. alm: score of the ALOM membrane spanning region prediction program
iv. mit: score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins
v. erl: presence of “HDEL” substring (thought to act as a signal for retention in the endoplasmic reticulum lumen).

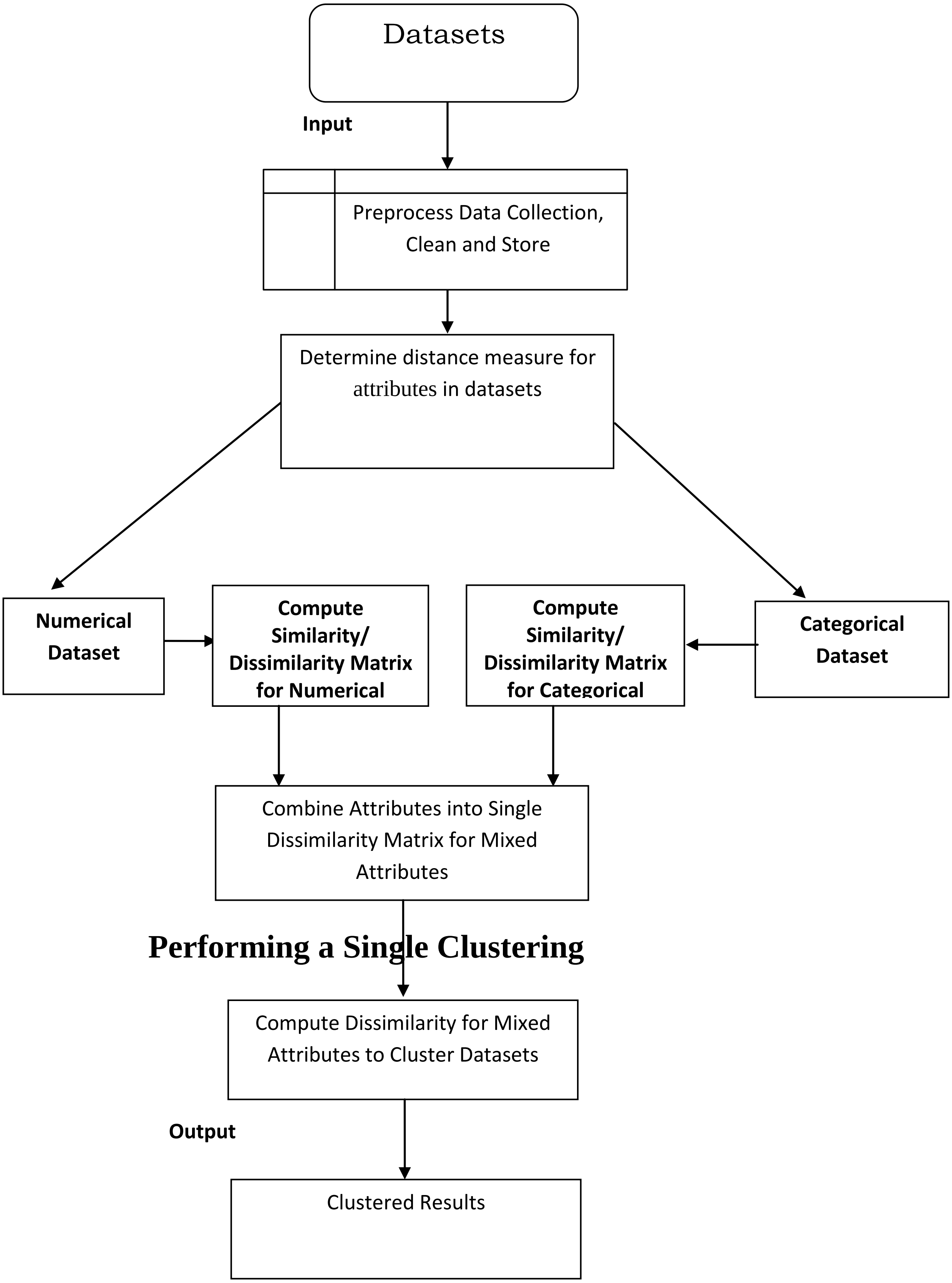

Figure 1: Proposed Hybrid Architecture

Table 1: Mixed datasets from soybean and yeast combined into a single table.

Sequence Name	Mcg	Gvh	Alm	Mit	Erl	Soybean_Name	Date	Plant stand	Precip	temp	hail
ADT1_ YEAST	0.58	0.61	0.47	0.13	0.50	Diaporthe-stem-canker	6	0	2	1	0
ADT2_ YEAST	0.43	0.67	0.48	0.27	0.50	Charcoal rot	6	0	0	2	0
ADT3_ YEAST	0.64	0.62	0.49	0.15	0.50	Rhizoctonia-root rot	1	1	2	0	0
AAR2_ YEAST	0.58	0.44	0.57	0.13	0.50	Phytophthora-rot	0	1	2	1	0
AATM_ YEAST	0.42	0.44	0.48	0.54	0.50	brown-stem-rot	4	0	0	1	0
AATC- YEAST	0.51	0.40	0.56	0.17	0.50	Porodery-mildew	5	0	0	1	1

In the other hand, the Sequence Name represents the class of soybean: accession member for the SWISS-PROT database, while the attributes are soybean-name, date, plant-stand, precip, temp, hail

where

i. Soybean_ name is the class of soybean in use. There are 19 classes from the original dataset but only 6 are being used to show how the dissimilarity is being calculated.
ii.
iii. Date: April, May, June, July, August, September, October
iv. Plant stand: normal, It- normal
v. Precip: It- norm, norm, gt- norm
vi. Temp: It- norm, norm, gt- norm
vii. Hail: yes, no

Therefore, the dissimilarity matrix for the mixed dataset is:

$$d(i,j) = \begin{bmatrix} 0 & & & & & \\ 0.40 & 0 & & & & \\ 0.41 & 0.61 & 0 & & & \\ 0.55 & 0.77 & 0.46 & 0 & & \\ 0.48 & 0.32 & 0.64 & 0.69 & 0 & \\ 0.52 & 0.51 & 0.69 & 0.46 & 0.45 & 0 \end{bmatrix}$$

where

d(i,j) is the measured dissimilarity or difference between objects i and j.

In this case, the lower triangular matrix can be interpreted as follows:

$$d(5,2) \to similar$$

$$d(4,2) \to dissimilar$$

The final dissimilarity matrix formed for the mixed dataset is in a Lower Triangular Matrix (LTM), which is the predictive model for identifying attributes (or objects) that are similar or dissimilar. The new algorithm has been implemented using JAVA and MATLAB 7.7 (R2008b).

Structured System Analysis and Design Methodology (SSADM) were used in this approach and two datasets obtained from UCI repository of machine learning databases (Asuncion et al. 2013) were used to demonstrate the way the clusters are formed.

Comparative Analysis of Frequency Distribution and Relative Cumulative Frequency (rcf)

Table 2 shows the data points extracted from yeast and soybean datasets. The relative cumulative frequency (rcf) computed demonstrates how the distribution could impact the way the clusters are formed as shown in table 3. The comparative analysis of frequency distribution and relative cumulative frequency are shown in figures 2 and 3 respectively.

The class interval was calculated using Sturges rule (Sturges, 1926). Sturges rule is a well-known statistical technique for determining class interval in a distribution. This rule is:

$$i = \frac{Range}{1+3.322*\log_{10} N}\text{...........equation (4)}$$

where Range = Highest number – lowest Number, N = total number of observation and

i = Class interval of the distribution.

From the frequency graph, it can be seen that the distribution is uneven. Most of the data points were found to be within the first few class intervals but as it moves to the right, there is a very sharp decline in the number of data points. On the contrary, the relative cumulative frequency graph demonstrates a more even distribution as the data points are spread out across the plot. This has ramifications to the clustering model in the sense that the data points are evenly spread and this reduces the risk of producing bad clusters. Table 3 shows the distribution.

RESULTS

The results in figure 6, for the value of k for instance, show how each cluster: cluster1, cluster 2,..........., cluster 8 contains attribute values that are similar to each other and those in other clusters are dissimilar to the other groups. From the results in appendix A, the braces clearly indicate the separate clusters grouped within an array. The diaporthe_stem_canker with [{6.0, 0.0, 2.0, 1.0}] are similar with diaporthe_stem_canker of [{5.0, 0.0, 2.1, 1.0}], which are closely related with charcoal_rot of [{6.0, 0.0, 0.0, 1.0}] and [{6.0, 0.0, 0.0, 2.0}] for soybean. In the same manner, yeast clusters of MIT[{0.58, 0.61, 0.47, 0.13}] are similar with MIT[{0.58, 0.44, 0.57, 0.13}]. Also the NUC[{0.42, 0.44, 0.48, 0.54}] are closely related with NUC [{0.42, 0.37, 0.59, 0.20}]. Finally, the CYT [{0.50, 0.54, 0.48, 0.65}] and CYT [{0.43, 0.39, 0.54, 0.21}] also share common features.

Table 2: Table Showing the Data Points from the Mixed Dataset.

0.58	0.43	0.64	0.58	0.42	0.51	0.47	0.48	0.49
0.57	0.48	0.56	0.13	0.27	0.15	0.13	0.54	0.17
0.50	0.50	0.50	0.50	0.50	0.50	6	6	1
0	4	5	0	0	1	1	0	0
2	0	2	2	0	0	1	2	0
1	1	1	0	0	0	0	0	1

Table 3: Computation of frequency and relative cumulative frequency-distribution

Class Interval	**Frequency**	**Cumulative Frequency**	**Relative Frequency**	**Relative Cummulative Frequency**
0 - 0.5	25	25	0.462962963	0.462962963
0.5 - 1.0	13	38	0.240740741	0.703703704
1.0 - 1.5	8	46	0.148148148	0.851851852
1.5 - 2.0	0	46	0	0.851851852
2.0 - 2.5	4	50	0.074074074	0.925925926
2.5 - 3.0	0	50	0	0.925925926
3.0 - 3.5	0	50	0	0.925925926
3.5 - 4.0	0	50	0	0.925925926
4.0 - 4.5	1	51	0.018518519	0.944444444
4.5 - 5.0	0	51	0	0.944444444
5.0 - 5.5	1	52	0.018518519	0.962962963
5.5 - 6.0	0	52	0	0.962962963
6.0 - 6.5	2	54	0.037037037	1
Total =	**54**			

The results show that the clusters formed share common characteristics and belong to one region. Since such clusters (or groups) identified are said to share some similarities, there is likelihood that they could also share in both pests problems and the equivalent treatments. It is deduced that what works for one cluster, may not work for another irrespective of the initial separate attributes of soybean and yeast datasets considered in table 2.

Appendix 1: JAVA output results of clustered soybean and yeast datasets

```
<*>><*><*><*><*>><*><*><*><*><*><*><*><*>Cluster8<*><*>><*><*><*><*>>>>>

[{[4.0, 0.0, 0.0, 1.0];charcoal-rot}]

[{[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0,
2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]

[{[5.0, 0.0, 0.0, 2.0];charcoal-rot}, {[5.0, 0.0, 0.0, 2.0];charcoal-rot}]

[{[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0,
0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0,
2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0,
1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]

[{[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0,
0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0,
2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]

[{[3.0, 0.0, 0.0, 1.0];charcoal-rot}]

[{[6.0, 0.0, 0.0, 1.0];charcoal-rot}, {[6.0, 0.0, 0.0, 1.0];charcoal-rot}]

[{[0.58, 0.61, 0.47, 0.13];MIT}, {[0.43, 0.67, 0.48, 0.27];MIT}, {[0.64, 0.62, 0.49, 0.15];MIT},
{[0.58, 0.44, 0.57, 0.13];NUC}, {[0.42, 0.44, 0.48, 0.54];MIT}, {[0.51, 0.4, 0.56, 0.17];CYT},
{[0.5, 0.54, 0.48, 0.65];MIT}, {[0.48, 0.45, 0.59, 0.2];NUC}, {[0.55, 0.5, 0.66, 0.36];MIT},
{[0.4, 0.39, 0.6, 0.15];CYT}, {[0.43, 0.39, 0.54, 0.21];NUC}, {[0.42, 0.37, 0.59, 0.2];NUC}]

[{[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0,
0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0,
2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0,
1.0];diaporthe-stem-canker}]

[{[6.0, 0.0, 0.0, 2.0]; charcoal-rot}, {[6.0, 0.0, 0.0, 2.0]; charcoal-rot}, {[6.0, 0.0, 0.0,
2.0];charcoal-rot}]
```

DISCUSSION

The figures 4 and 5 are screenshots of MATLAB simulation performed on soybean and yeast datasets and the equivalent JAVA sample output is also shown in figure 6. The clustering was performed at random values of k. From the output of the clusters, it is observed that the results consist of a combination of the attributes of yeast and soybean in one cluster. For the purpose of clarity, an extract of JAVA output has been appended as Appendix A. The braces indicate the separate clusters grouped within an array. The use of the new hybrid, Mixk-meansXFon, clustering approach could serve as means to mitigate the spread of pest by first identifying the region in the field in which the crops are affected and then performing the clustering algorithm to cluster and separate affected crops. The clustering algorithm helps to identify the crops that are affected by the pest and hence targeted treatment can then be pursued. This prevents the farmer from carrying out full scale treatment of the entire crops by allowing him focus on the affected areas (or region). This approach saves cost and time, which is also in line with the IPM program in the agricultural sectors. This could be beneficial to the Nigerian farmers and Ministry of Agriculture, particularly the departments that handle crop scouting and those that determine the economic thresholds (or action threshold) of pests and disease control.

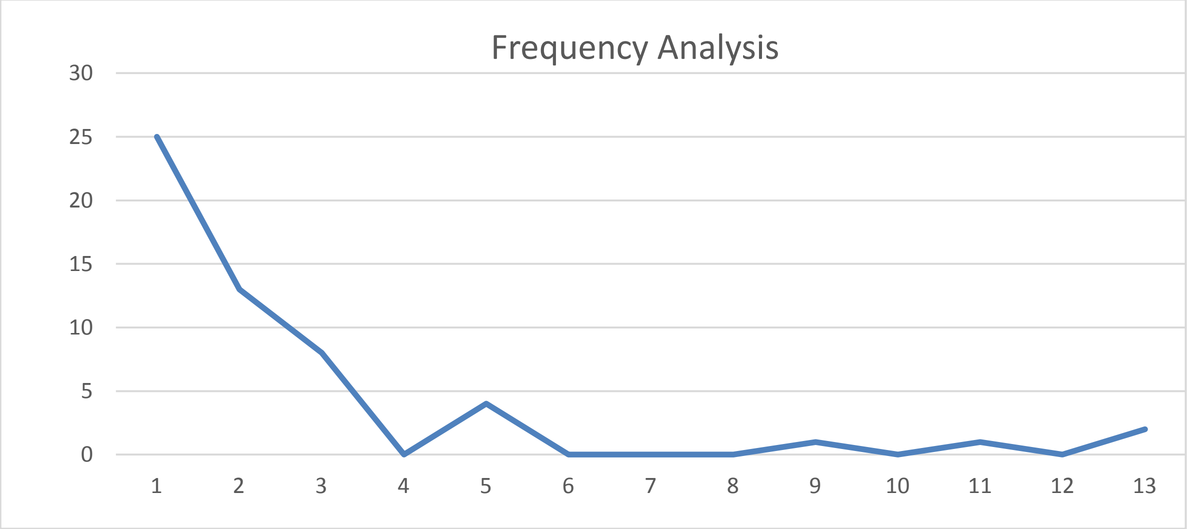

Figure 2: Frequency Analysis graph

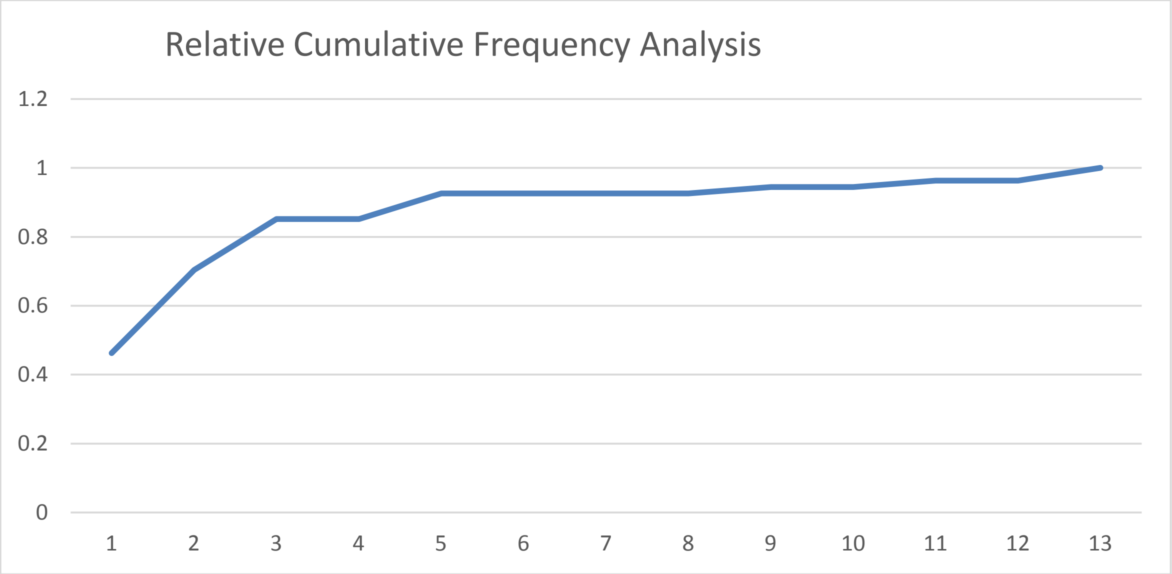

Figure 3: Relative Cumulative Frequency Analysis graph

Scientia Africana, Vol. 13 (No.2), December 2014. Pp225-241
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria ISSN 1118 - 1931

SNAPSHOT OF MATLAB SIMULATION OF SOYBEAN

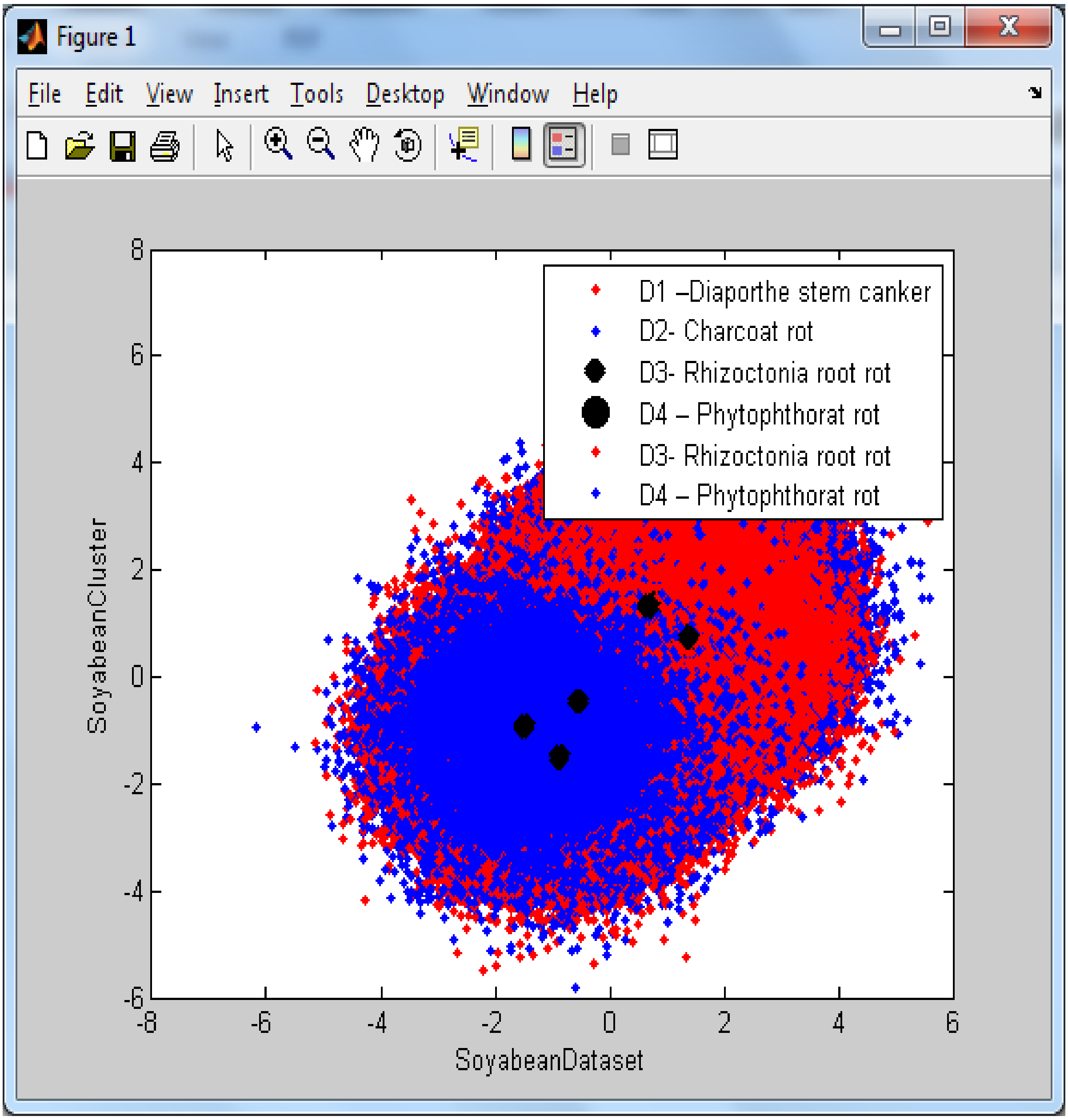

Figure 4: MATLAB simulation output of clusters of soybean.

SNAPSHOT OF MATLAB SIMULATION OF YEAST

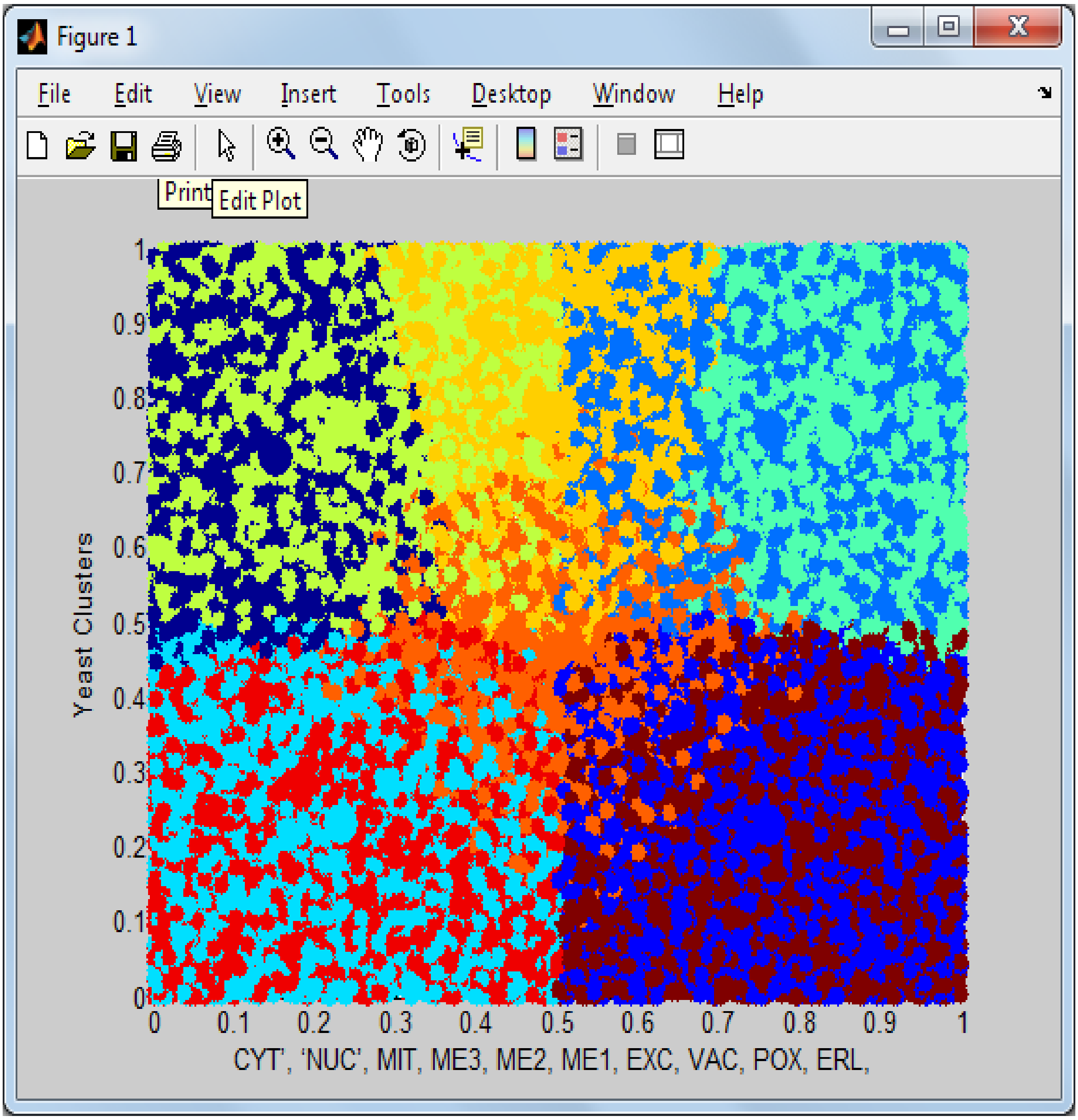

Figure 5: MATLAB simulation output of clusters of yeast.

Scientia Africana, Vol. 13 (No.2), December 2014. Pp225-241
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria ISSN 1118 - 1931

SNAPSHOT OF MIXED DATA CLUSTERING

CHOOSE A DATASET: Mixed Dataset

RANDOM GENERATED VALUE OF K: 8

SUM OF SQUARE ERROR: 0.06

```
[{[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]
[{[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-st
em-canker}]
[{[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]
[{[0.58, 0.61, 0.47, 0.13];MIT}, {[0.43, 0.67, 0.48, 0.27];MIT}, {[0.64, 0.62, 0.49, 0.15];MIT}, {[0.58, 0.44, 0.57, 0.13];NUC}, {[0.42, 0.44, 0.48, 0.54];MIT}, {[0.51, 0.4, 0.56, 0.17
];CYT}, {[0.5, 0.54, 0.48, 0.65];MIT}, {[0.48, 0.45, 0.59, 0.2];NUC}, {[0.55, 0.5, 0.66, 0.36];MIT}, {[0.4, 0.39, 0.6, 0.15];CYT}, {[0.43, 0.39, 0.54, 0.21];NUC}, {[0.42, 0.37, 0.59,
0.2];NUC}]
[{[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}]
[{[6.0, 0.0, 0.0, 2.0];charcoal-rot}, {[6.0, 0.0, 0.0, 1.0];charcoal-rot}, {[6.0, 0.0, 0.0, 2.0];charcoal-rot}, {[5.0, 0.0, 0.0, 2.0];charcoal-rot}, {[6.0, 0.0, 0.0, 1.0];charcoal-rot}, {[6.0,
0.0, 0.0, 2.0];charcoal-rot}, {[5.0, 0.0, 0.0, 2.0];charcoal-rot}]
[{[3.0, 0.0, 0.0, 1.0];charcoal-rot}]


<*>><*><*><*>><*><*><*><*><*><*><*><*>Cluster8<*><*>><*><*><*>><*><*><*><*><*><*><*><*><*>>>>
[{[4.0, 0.0, 0.0, 1.0];charcoal-rot}]
[{[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[3.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]
[{[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[6.0, 0.0, 2.0, 1.0];diaporthe-st
em-canker}]
[{[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[4.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}]
[{[0.58, 0.61, 0.47, 0.13];MIT}, {[0.43, 0.67, 0.48, 0.27];MIT}, {[0.64, 0.62, 0.49, 0.15];MIT}, {[0.58, 0.44, 0.57, 0.13];NUC}, {[0.42, 0.44, 0.48, 0.54];MIT}, {[0.51, 0.4, 0.56, 0.17
];CYT}, {[0.5, 0.54, 0.48, 0.65];MIT}, {[0.48, 0.45, 0.59, 0.2];NUC}, {[0.55, 0.5, 0.66, 0.36];MIT}, {[0.4, 0.39, 0.6, 0.15];CYT}, {[0.43, 0.39, 0.54, 0.21];NUC}, {[0.42, 0.37, 0.59,
0.2];NUC}]
[{[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-canker}, {[5.0, 0.0, 2.0, 1.0];diaporthe-stem-ca
nker}]
[{[6.0, 0.0, 0.0, 2.0];charcoal-rot}, {[6.0, 0.0, 0.0, 1.0];charcoal-rot}, {[6.0, 0.0, 0.0, 2.0];charcoal-rot}, {[5.0, 0.0, 0.0, 2.0];charcoal-rot}, {[6.0, 0.0, 0.0, 1.0];charcoal-rot}, {[6.0,
0.0, 0.0, 2.0];charcoal-rot}, {[5.0, 0.0, 0.0, 2.0];charcoal-rot}]
[{[3.0, 0.0, 0.0, 1.0];charcoal-rot}]
```

Figure 6: Shows the clustering of mixed datasets of soybean and yeast using new hybrid algorithm

The use of mixed dataset for this clustering was necessitated by real world scenarios in which two or more datasets are interacting with each other. Crops mostly grouped in any field are susceptible to being attacked by pest. There has to be different measures to curtail the spread of these pests within the crops. The new hybrid, Mixk-meansXFon, clustering approach has been developed and integrated into the IPM program to serve as means to mitigate the spread of pest by first identifying the region in the field in which the crops are affected and then performing the clustering algorithm to cluster and separate affected crops.

The clustering algorithm helps to identify the crops that are affected by the pest and hence targeted treatment can then be pursued. This prevents the farmer from carrying out full scale treatment of the entire crops by allowing him focus on affected areas. This approach saves cost and time, which is also in line with the IPM program in agricultural sectors. This could be beneficial to the Nigerian farmers and Ministry of Agriculture, particularly the departments that are involved with crop scouting and those that determine the economic thresholds (or action threshold) of pests and disease control. In this work, the new hybrid clustering algorithm has been proposed on the management of crop pests. In the control of crop diseases, once a pest has been found to exceed the economic threshold, the farmer must determine the best and least expensive way to prevent impending losses. If a pesticide treatment is needed, the rate and method of pesticide application must be customized. This is probably the surest way to save money.

Future work

The hybrid algorithm is only designed primarily to help scouts and growers learn how to check for crop pests and to know when treatment is economically justified. The pesticide guide is not included in the recommendation, which is a major limitation in this work. There is still lot more open problems especially integrating the new algorithm, MixK-meansXFon, into the integrated pest management (IPM) programs and applying this method on crops (maize, cassava, yam etc.) grown around here in Nigeria.

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 Sc. Information Tech. and Management, 1(1-2).

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

Babajide, O.T., Akinlabi, B.H. and Tijani, A.A.(2012), Agricultural Resource and Economic Growth in Nigeria, European Scientific Journal, 105

Dubey, S.R., Dixit, P., Singh, N. and Gupta, J.P. (2013), Infected Fruit Part Detection Using K-means Clustering Segmentation Technique, International Journal of Artificial Intelligence and Interractive Multimedia, 2 (2), DOI:10-9781/ijimai.229.

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). 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.

Kleinberg, J. (2002). An Impossibility Theorem for Clustering Nips 15, MIT Press, 446-453.

Kolhe, S.R. and Zinjore, B.S. (2010), Clustering Iris Data Using Supervised and Unsupervised Learning, International Journal of Computer Science and Application, ISSN: 0974-0767.

Kogan, M. (1998) Integrated Pest Management; Historical Perspectives and Contemprary Development, Annual review of Entomology, 43, 243-278.

Ramesh, V., Ramar, K.and Babu, S. (2013), Parallel K-means Algorithm on Agricultural Database, IJCS International Journal of Computer Science Issues, 10, Issue 1, 1, ISSN (print) 1694-0784), www.IJCSI.org

Sturges H.A (1926). "The Choice of a Class Interval". Journal of the American Statistical Association.

Soybean Pest Management (2000). Retrieved on 5th May, 2013 from http://ipm.ncsu.edu/soybeans/Scouting_Soybeans/soybean_pest_management.html,

Soybean Field Guide (2001) 2nd Edition, 23-24.