CLASSIFICATION OF INFECTIOUS DISEASES VIA HYBRID *K*-MEANS CLUSTERING TECHNIQUE

[1,2]*DAUDA USMAN AND [1]ISMAIL BIN MOHAMAD
[1]Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia

[2]Department of Mathematics and Computer Science, Faculty of
Natural and Applied Sciences
Umaru Musa Yar'adua University, Katsina
P.M.B 2218 Katsina-Nigeria

[1]dauusman@gmail.com, [2]ismailm@utm.my

*Corresponding author

Abstract. Identifying groups of objects that are similar to each other but different from individuals in other groups can be intellectually satisfying, profitable, or sometimes both. *K*-means clustering is one of the well known partitioning algorithms. But basic *K*-means method is insufficient to extract meaningful information and its output is very conscious to initial positions of cluster centers. In this paper, data of infectious diseases were analyzed with the hybrid *K*-means clustering technique. This method is developed to preprocess the dataset that will be used in the *K*-means clustering problems. Specifically, it performs *K*-means clustering on preprocessed dataset instead of raw dataset to remove the impact of irrelevant features and selection of good initial centers. The experimental results revealed that all the three water related diseases are grouped together in one cluster for both KGHK and FMCK data sets. They also show the high prevalence compared to airborne particle related diseases in the other group. The study concludes that *K*-means clustering method provides a suitable tool for assessing the level of infectious diseases.

Keywords data standardization; center initialization; cluster analysis; infectious diseases; *k*-means algorithm; principal component analysis.

1.0 INTRODUCTION

Diseases affecting humans are caused by infection such as leprosy, chicken pox and typhoid fever [1]. The aetiology of some of these diseases is induced by environmental factors. Infection differs from other diseases in a number of aspects. The most important is that it is caused by living microorganisms which can usually be identified, thus establishing the aetiology early in the illness. Many of these organisms, including all bacteria, are sensitive to antibiotics and most infections are potentially curable, unlike many non-infectious diseases which are degenerative and frequently become chronic. Communicability is another factor which differentiates infectious from non-infectious diseases. Transmission of pathogenic organisms to other people, directly or indirectly, may lead to an epidemic.

Many infections are preventable by hygienic measures, vaccines or judicious use of drugs (chemoprophylaxis) [2].

Statistical analysis has become an important tool in understanding the dynamics of diseases transmission and in decision making processes regarding intervention programs for disease control. Recently, clustering has been applied to many areas of health psychology, including the promotion and maintenance of health, improvement to the health care system, and prevention of illness and disability [3]. *K*-means clustering algorithm is one of the most popular methods for clustering multivariate observations [4]. It is a system ordinarily used to directly segment sets of data into *k* groups. *K*-means algorithm generates a fast and efficient solution. The basic *K*-means algorithm works with the objective to minimize the mean square distance from each data point to its nearest center.

In the creation of this *K*-means clustering algorithm two main issues are prominent, these are: the optimal number of clusters and the cluster center points. In most cases, the number of clusters is given, thus leaving the challenge where to put the cluster centers so that scattered points can be grouped properly and to avoid its convergence to a local minimum of the objective function. Furthermore, the random initialization results in different total SSEs value from several runs of the *K*-means. This makes the result from the algorithm of the *K*-means to depend greatly on the initial selection of the cluster centers which must be known and fixed beforehand. Therefore, the choice of proper initial centroids is pertinent to the basic *K*-means procedure.

Though, many attempts were made by different researchers to improve the effectiveness and efficiency of the basic *K*-means algorithm. Fahim [5] proposed a method to select a good initial solution by partitioning data set into blocks and applying *K*-means to each block. Tajunisha and Saravanan [6] proposed a method to improve the performance of the *K*-means algorithm, using principal component analysis (PCA) for dimension reduction and to find the initial centroid for *K*-means. The method partitioned the data set into *K* sets and the median of each set were used as initial cluster centers and then assign each data point to its nearest cluster centroid. Heuristic approach was also used to reduce the number of distance calculation in the standard *K*-means algorithm to assign the data point to the cluster. But the time complexity for the two methods above is slightly more.

Mohammed and Wesam [7] proposed a visual clustering framework using C++ Builder 2009, for initialization of the *K*-means clustering algorithm. The method generates *K* points using semi random technique. It makes the diagonal of the data as a starting line and selects the points randomly around it. But this method did not suggest any improvement to the time complexity for *K*-means algorithm. Al-Shboul and Myaeng [8] proposed two algorithms to solve the initialization problem and these are: genetic algorithm initializes *K*-means (GAIK) and *K*-means initializes the genetic algorithm (KIGA). A new model was built that enhances the quality of the clustering (reduces the future error), and carried out a comparative analysis amongst GAIK, KIGA, genetic-based clustering algorithm (GCA) indicating that KIGA is best compared to others. This was proven from the outcomes on all

datasets that KIGA can achieve higher clustering accuracy when compared with other algorithms.

Marghny [9] proposed an improved genetic *K*-means algorithm (IGK) to handle large scale data, which can select an initial clustering center using genetic algorithms (GAs). The Procedure finds an approximate solution to such problems through the application of the principles of evolutionary biology. The dimension of the data was reduced via principal component analysis and employed a medoid in order to diminish the sensitivity of initial point choice which is the most centrally located object in a cluster, to get good initial centroids. The computational cost of this method is also quite high and there is every possibility of choosing bad initial centroids, such as outliers that may lead to the generation of singletons, or perhaps obtaining a set of initial centroids that are too close to one another.

Al Hasan [10] use a local outlier factor (LOF) to avoid selecting outlier points as centers. In the iteration $i(i \epsilon \{1, 2, \dots, k\}$, the method first sorts the data points in descending order of their minimum distance to the previously chosen centers. After that it rotates the points in sorted order and selects the first point that has a LOF value close to 1 as the *ith* center. The challenge with this technique is that the computational cost is dominated by its complexity of sorting the data in descending order, which is $0(NlogN)$.and there is no assurance that the accuracy of the final clusters selected is guaranteed. All the algorithms presented in this literature are quite complex and used the *K*-means algorithm as part of their algorithm, which still need to use the random method for cluster center initialization.

As fallout from the above, a new technique of centers initialization for *K*-means clustering is required to make the algorithm more effective and efficient. Hence this paper focuses on an alternative way of handling the *K*-means clustering technique and classification of infectious diseases data sets by data pre-processing and center initialization method.

2.0 MATERIALS AND METHODS

The technique is a hybrid of the basic *K*-means method. First, *z*-score standardization is used to scale the data to fall within a specified range of values, so that any variable with larger domain will not dominate the variable with smaller domain. Second, principal component analysis using the singular value decomposition is used to obtain a reduced data set containing possibly uncorrelated variables. Third, the resulting reduced data set is applied to the *K*-means clustering algorithm. All the analysis is carried out using MATLAB M-File 7.6 (R2008a) package. The steps for the technique are as follows:

Step 1

Consider the z-score method of data standardization. For convenience, let $\boldsymbol{X} = (\boldsymbol{X}_1^{'}, \boldsymbol{X}_2^{'}, \dots, \boldsymbol{X}_n^{'})$ is the d-dimensional raw data set. This results in an $n \times d$ data matrix given by:

$$\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \ldots, \boldsymbol{X}_n') = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

The z-score of $\boldsymbol{X}_i$ is given as

$$Z(\boldsymbol{X}_i) = \frac{x_i - \bar{x}_i}{S_{x_i}} \quad (1)$$

where x_i, is the raw scores to be standardized. $\bar{x}_i$ and S_{x_i} are the sample mean and sample standard deviation respectively.

Step 2

In step 2, we proceed to find the principal components of the standardized dataset in Equation 1. After finding the principal components of the data, a reduced projected dataset will also be obtained by multiplying the original and principal components data matrices and retaining 80 percent of the total variance. Consider the standardized data matrix $\boldsymbol{X}$, with zero mean and variance 1. Mathematically, the transformation is defined by a set of p-dimensional vectors of loadings (loadings here means the weight by which each standardized original variable should be multiplied to get the component score).

$\boldsymbol{w}_h = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_p)_h$ that map each row vector $\boldsymbol{X}_i$ of $\boldsymbol{X}$ to a new vector of principal component scores
$\boldsymbol{t}_{(i)} = (t_{(1)}, t_{(2)}, \ldots, t_{(p)})_{(i)}$, given by

$$\boldsymbol{t}_{h(i)} = \boldsymbol{X}_{(i)} . \boldsymbol{w}_{(h)} \quad (2)$$

in such a way that the individual variables of $\boldsymbol{t}$ considered over the data set successively inherit the maximum possible variance from $\boldsymbol{X}$, with each loading vector $\boldsymbol{w}$ constrained to be a unit vector.

- The 1st component

The 1st loading vector $\boldsymbol{w}_{(1)}$ thus has to satisfy

$$arg \max_{\|w\|=1} \{\Sigma_i (t_1)^2_{(i)}\} = arg \max_{\|w\|=1} \{\Sigma_i (\boldsymbol{X}_i . \boldsymbol{w})^2\} \quad (3)$$

Consistently, writing this in a matrix structure will give

$$\boldsymbol{w}_{(1)} = arg \max_{\|w\|=1} \{\|\boldsymbol{w}\|^2\} = arg \max_{\|w\|=1} \{\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}\} \quad (4)$$

As $\boldsymbol{w}_{(1)}$ is a unit vector which is equally satisfied

$$\boldsymbol{w}_{(1)} = argmax \left\{ \frac{\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \right\} \quad (5)$$

A standard result for a symmetric matrix, $\boldsymbol{X}^T \boldsymbol{X}$ is that the quotient's maximum possible value is the largest eigenvalue of the matrix. This results when, $\boldsymbol{w}$ is the eigenvector that

corresponds to it. With $\boldsymbol{w}_{(1)}$ as our first component of a data vector $\boldsymbol{X}_i$. This can also be given as a score of $t_{1(i)} = \boldsymbol{X}_i.\boldsymbol{w}_i$ in the transformed co-ordinates, or as the corresponding vector in the original variables, $\{\boldsymbol{X}_i.\boldsymbol{w}_{(1)}\}\boldsymbol{w}_1$.

- For further components

The hth component can be obtained by taking the difference of the first $h-1$ principal components from $\boldsymbol{X}$.

$$\widehat{\boldsymbol{X}}_{h-1} = \boldsymbol{X} - \sum_{s=1}^{h-1} \boldsymbol{X}\boldsymbol{w}_{(s)}\boldsymbol{w}_{(s)}^{'} \quad (6)$$

The loading vector is one that extracts the maximum variance from this new data matrix

$$\boldsymbol{w}_{(h)} = arg \max_{\|\boldsymbol{w}\|=1} \left\{\left\|\widehat{\boldsymbol{X}}_{h-1}\boldsymbol{w}\right\|^2\right\} = argmax \left\{\frac{\boldsymbol{w}^{'}\widehat{\boldsymbol{X}}_{h-1}^{'}\widehat{\boldsymbol{X}}_{h-1}\boldsymbol{w}}{\boldsymbol{w}^{'}\boldsymbol{w}}\right\} \quad (7)$$

Step 3

Given a set of observations, $\widehat{\boldsymbol{X}} = (\widehat{\boldsymbol{X}}_1', \widehat{\boldsymbol{X}}_2', \ldots, \widehat{\boldsymbol{X}}_n')$ where each observation is a p-dimensional real vector, to partition the n observations into k sets $(k \leq n)$, $G = (g_1, g_2, \ldots, g_k)$ compute:

$$d_{euclidean}(\boldsymbol{XY}) = \sqrt{(x_i - y_i)^2} = [(x-y)(x-y)^{'}]^{\frac{1}{2}} \quad (8)$$

The algorithm proceeds by alternating between two steps

$$G_i = \{x_p : \|x_p - \mu_i\|^2 \leq \|x_p - \mu_j\|^2 \; \forall \, j, 1 \leq j \leq k\} \quad (9)$$

where, each x_p is assign to exactly one G. Then update the process by calculating the new centers in the new clusters. The algorithm converges when this assignment no longer changes. Then calculate the total sum of squares error (SSE) that is:

$$SSE = arg_g min \sum_{i=1}^{k} \sum_{x_j \in c_j} \|x_j - \mu_j\|^2 \quad (10)$$

where, $\mu_i = \frac{1}{n}\sum_{x_i \in c_j} \boldsymbol{X}_i$ denotes the centroid of a cluster c_j and n denotes the number of instances in c_j.

3.0 RESULTS AND DISCUSSIONS

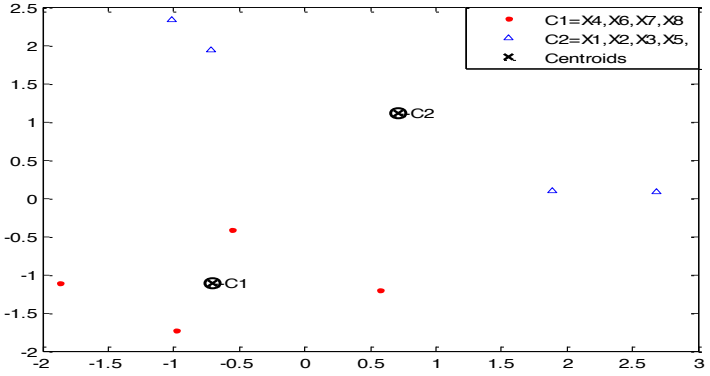
A life data of infectious diseases is used in this paper. These data sets are secondary data obtained from Katsina General Hospital, Katsina (KGHK) and Federal Medical Centre, Katsina (FMCK) consisting of eight variables and sample size 36. The eight variables are Malaria, Typhoid fever, Cholera, Measles, Chicken pox, Tuberculosis, Tetanus and Leprosy which are denoted by X_1 to X_8 respectively. The first three variables denoted by X_1, X_2 and X_3 are water related diseases while the variables denoted by X_4, X_5, X_6, X_7 and X_8 are airborne particle diseases. These diseases X_1 to X_8 are recorded over 36 months (three years) from January, 2011 to December, 2013 and are referred to as sample size 36. These numbers represents the number of occurrences of each disease within a month.

The analysis and cluster formations plots were carried out using the hybrid *K*-mean clustering method. Two different cluster formations are constructed for Katsina General Hospital Katsina (KGHK) data and Federal Medical Centre Katsina (FMCK) data. These are presented in Figure 3.1 and Figure 3.2 respectively. The infectious diseases with similar degree of prevalence are grouped together into a cluster. This is done using squared Euclidean distance measure as in [11] which says a distance is used to measure the similarity among objects, in such a way that more similar objects have lower dissimilarity value. The distances, cases number which corresponds to their clusters are shown in Table 3.1 for KGHK data and FMCK data respectively.

Table 3.1: Cluster cases number and their distances of KGHK data

	KGHK		**FMCK**	
Case Number	Cluster	Distance	Cluster	Distance
X_1	2	5.854	1	5.183
X_2	2	6.216	1	5.025
X_3	2	5.134	1	4.327
X_4	1	3.277	2	2.599
X_5	2	4.879	2	3.621
X_6	1	3.375	2	2.706
X_7	1	0.438	2	0.526
X_8	1	0.194	2	0.381

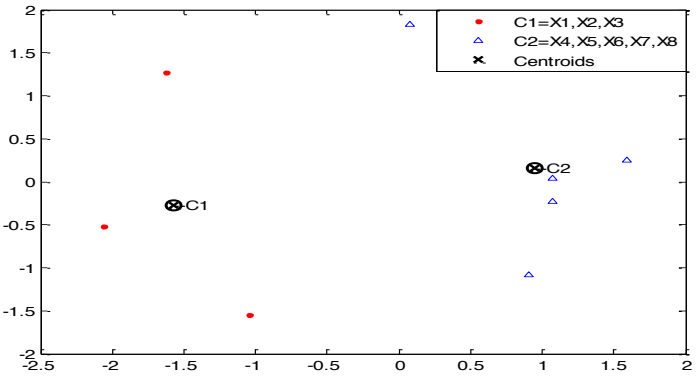
Table 3.1 show the cases number and distances obtained which corresponds to their clusters using the preprocessed Katsina General Hospital, Katsina (KGHK) data and Federal Medical Centre, Katsina (FMCK) data respectively. The cluster formations for these are presented in Figure 3.1 and Figure 3.2 respectively.



Malaria=X_1, Typhoid fever=X_2, Cholera=X_3, Measles=X_4, Chicken pox=X_5, Tuberculosis=X_6, Tetanus=X_7 and Leprosy=X_8

Figure 3.1 Classification of KGHK data into two clusters

Figure 3.1 shows the classification of infectious diseases found with KGHK data. The classification was achieved by clustering the preprocessed data into clusters with their patterns using the hybrid *K*-means method. It can be observed that the infectious diseases with similar degree of prevalence are grouped together in clusters. The four airborne particle diseases are grouped into cluster 1 while all three water related diseases and one airborne particle disease are grouped into cluster 2.



Malaria=X_1, Typhoid fever=X_2, Cholera=X_3, Measles=X_4, Chicken pox=X_5, Tuberculosis=X_6, Tetanus=X_7 and Leprosy=X_8

Figure 3.2 Classification of FMCK data into two clusters

Figure 3.2 shows the classification of infectious diseases found with FMCK data into clusters and the pattern of these diseases using the hybrid *K*-means method. It can be seen that the infectious diseases with similar degree of prevalence are grouped together in clusters. The three water related diseases are grouped into cluster 1 and the five airborne related diseases in cluster 2.

It can be observed from Figure 3.1 that four airborne particle diseases out of five namely measles, tuberculosis, tetanus and leprosy are grouped into cluster 1 as they show less prevalence compared to those in cluster 2 and all are airborne particles diseases. This can be observed by their small distances from the cluster centre as shown in Table 3.1. Cluster 2 comprises of three water related diseases and one airborne particle disease namely malaria, typhoid fever, cholera and chickenpox are grouped into cluster 2 and they are the most prevalent diseases. This can be seen from the large distances of these diseases in relation to the cluster centre as shown in Table 1. The cluster centers are marked c_1 and c_2 for cluster 1 and cluster 2 respectively.

In Figure 3.2, it can be observed that all the three water related diseases namely malaria, typhoid fever and cholera are grouped into cluster 1. This indicates that they are the most prevalent diseases. This can be seen from the large distances of these diseases in relation to the cluster centre as can be seen in Table 3.1. Cluster 2 comprises of measles, chickenpox, tuberculosis, tetanus, and leprosy as they show less prevalence compared to those in cluster 1 and these are airborne particles diseases. This can be observed by their small distances from the cluster centre as can be seen in Table 3.1. The cluster centers are marked c_1 and c_2 for cluster 1 and cluster 2 respectively.

4.0 CONCLUSION

From the characteristics of the diseases they can be divided into three groups namely airborne particles diseases, water related diseases and the frequency of occurrence of the disease. The bacterium that causes most of the diseases in cluster 2 in Figure 3.1 and cluster 1 in Figure 3.2 are spread mostly through poor hygiene habits like contaminated drinking water and poor public sanitary conditions, and sometimes by flying insects that feeding on excreta. Those found in cluster 1 in Figure 3.1 and cluster 2 in Figure 3.2 are caused by airborne particles or contagious bacteria related and can be spread through droplet transmission from the nose, throat, and mouth of someone who is infected with the virus.

As such the most important components in controlling the spread of the diseases in cluster 2 in Figure 3.1 and cluster 1 in Figure 3.2 is to introduce public education campaigns; encourage citizens to wash their hands after defecating and before eating and to go for vaccination regularly. It is also, recommended that good and proper sanitation be observed; proper care be taken in food preparation and sanitary health workers should pay regular visits to homes and public eateries to supervise cleaning. In cluster 1 in Figure 3.1 and cluster 2 in Figure 3.2 the most importance measure is regular vaccination. These are some of the critical measures that can be taken to prevent the diseases in cluster 1. Lastly, the government and other authorities should discharge their responsibilities properly by providing good public drainage system, ensuring environmental protection and provision of good pipe borne water; these measures can help reduce the level of susceptibility of citizens to any form of disease.

REFERENCES

[1] Bloom, A. (1963). Toohey's Medicine for Nurses. Whittington London.

[2] Davidson, S. (2006). *Principle and Practice of Medicine, 20th Edition*. Edinburg, New York.

[3] Clatworthy, J., Buick, D., Hankins, M., Weinman, J. & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology*, 10(3):329–358.

[4] Tsai, C. Y. & Chiu, C. C. (2008). Developing a Feature Weight Self-Adjustment Mechanism for a *K*-means Clustering Algorithm. *Computational Statistics and Data Analysis*, 52:4658-4672.

[5] Fahim, A. M., Salem, A. M., Torkey, F. A., Saake, G. & Ramadan, M. A., (2009). An Efficient K-means with Good Initial Starting Points, *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*. 2(19):47-57.

[6] Tajunisha, N. and Saravanan, V. (2010). Performance analysis of *K*-means with Different Initialization Methods. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 1(4):44-52.

[7] Mohammed, E. and Wesam, M. A. (2012). Efficient and Fast Initialization Algorithm for *K*-means Clustering. *International Journal of Intelligent Systems and Applications*. 1:21-31.

[8] Al-Shboul, B. and Myaeng, S. (2009). Initializing *K*-means Using Genetic Algorithms. *World Academy of Science, Engineering and Technology*, 54:114-118.

[9] Marghny, M. H., Rasha M. A. & Ahmed, I. T. (2011). An Effective Evolutionary Clustering Algorithm: Hepatitis C Case Study. *International Journal of Computer Applications*. 34(6):01-06

[10] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M. (2009). Robust Partitional Clustering by Outlier and Density Insensitive Seeding. *Pattern Recognition Letters*. 30(11):994–1002.

[11] Doreswamy and Hemanth, K. S. (2012). A Novel Design Specification Distance (DSD) Based K-Mean Clustering Performance Evaluation on Engineering Materials' Database. *International Journal of Computer Applications*, 55(15): 26-33.