FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 2 No. 3, September, 2018, pp. 24 - 33

OPTIMIZATION OF K-MODE ALGORITHM FOR DATA MINING USING PARTICLE SWARM OPTIMIZATION

[*1]Obuandike, G. N., [2]John, A. and [3]Ismaila, I.

[1]Department of Computer Sciences and IT, Federal University Dutsin-Ma, Katsina state.
[2]Department of Computer Science, Federal University of Technology, Minna, Niger State
[3]Department of Cyber Security, Federal University of Technology, Minna, Niger State.

Corresponding Author's email: gobunadike@fudutsinma.edu.ng

ABSTRACT
K-mode is a popular data mining algorithm because of its effective performance in handling categorical data. It has a problem in its methodology in the area of choosing the initial cluster centers for its clustering tasks which usually affects its results. The research proposed a novel PSO K-mode algorithm called PSOKM to improve the performance of K-mode clustering algorithm using PSO. Fitness function was defined based on the structure of K-mode algorithm and weights; the cluster centroids were optimized using PSO. The initial cost for the PSO was taken from K-mode; the weights were picked at random and two centroids from each class were randomly picked. The research used University of California Irvine (UCI) data set and crime data to evaluate the performances of the PSOKM algorithms against conventional K-mode algorithms using metrics such as accuracy, time, sensitivity, specificity and ROC curve. Evaluation result reveals that the PSOKM improved the accuracy of K mode algorithm from 76% to 89.4% using the crime data. The reliability of the algorithms performance was also conducted using UCI data set and the results obtained were compared with the ones from other variant algorithms. The result revealed that the performance of PSOKM were better than that of the respective variants in most cases.

Keywords: Data Mining, Clustering, Particle Swarm Optimization, K-mode

INTRODUCTION

Data mining has gained acceptance as popular research area in computer science and other related areas. This is because it has application in different areas where it is been applied to uncover hidden patterns in data. It incorporates techniques from other fields like mathematics, statistics, machine learning, artificial intelligence and database (Jiawei, Micheline and Jian, 2012). Clustering of data is one of the many activities in data mining that is used to group data according to similarities (Chuang et al, 2012). K-mode algorithm is a popular algorithm that is applied in grouping categorical data but has a challenge in its methodology in selection of initial cluster centroids, double grouping of data items and premature convergence (Huang, 1997; Zhang et al, 2000; Huang, 2002). Many techniques have been applied to improve its performance in data clustering. The application of artificial intelligence techniques to improve algorithm performance is receiving attention in research in recent years (Ghorpade-Aher and Metre, 2014). Particle Swarm Optimization (PSO) is a subfield of swarm intelligence under Artificial Intelligence which studies the emergent collective intelligence of groups of simple agents. It applies the social behavior usually observed in birds flocks. There are other swarm intelligence techniques but PSO is faster, simpler and more efficient compared to other swarm intelligence techniques (Martens et al, 2011). PSO has been choosing as a technique to be used to improve the performance of k-mode algorithm. The rest of the paper is on the concept of k-mode algorithm; it's optimization with PSO technique and the discussion of the methodology used in the optimization and evaluation result.

K-MODE ALGORITHM

Majority of data that are gathered from real world activities are usually non numeric data (Zhao & Mei Lu, 2013). K-mode is famous for its ability to handle non numeric data effectively and can handle huge data set. It is an expansion of K-means in order to handle non numeric data. It was designed to handle non numeric data and it uses mode instead mean in its methodology. It applies comparison method in its methodology in handling categorical data by using mode in place of mean in order to reduce the cluster cost function (Huang and Ng, 2003).

This update allows K-mode work in a way similar to the workings of K-means, it replaces the Euclidean distance function of K-means by comparison method. It uses mode to calculate the cluster centres and updates mode value using the most repeated data item. There is always the challenge of picking the initial cluster centres which usually affects the result. This has been the major drawback of K-mode which many literatures have tried to address (Zhao & Mei Lu, 2013).

Supposed b and c are non-numeric dataset with M fields. The comparison method of two data item can be written as $d(b,c)$ between b and c and it taken to be the total number comparison value of the two corresponding data item. The smaller this value the more alike the two data items. The comparison method $d(b,c)$ can be written in mathematical format as:

$$d(b,c) = \sum_{i=1}^{m} d(b_i, c_i) \quad (1)$$

Where $d(b_i, c_i) = \begin{cases} 0 & (b_i = c_i) \\ 1 & (b_i \neq c_i) \end{cases}$ (2)

If z is vector of non-numeric set with attributes $x_1, x_2 \ldots x_m$. When plugged into equation 1 as comparison method for categorical data it will give a cost function represented as:

$$f(\emptyset) = \sum_{i=1}^{n} d(z_i, q_i) \quad (3)$$

Where z_i is the i[th] element and q_i is the nearest cluster centre to of z_i. Kmode algorithm minimizes cost function (Zhao & Mei Lu, 2013).

PSO TECHNIQUE AND CONCEPT

PSO is an acronym for Particle Swarm Optimization and is a branch of Swarm Intelligence in Artificial Intelligence, it is a new paradigm that has received wide spread attention in research (Satyobroto, 2011). PSO copies the behavior of birds that move together in a group in search of food or better opportunities. It gives better result in difficult problems and has few parameters to work with. It is fast and accurate in its methodology and this has promoted its general acceptance in optimization (Satyobroto, 2011; Martens et al, 2011).

PSO is usually made up of three basic features which are the particle, particle experiences and velocities. In a problem space where there may be more than one possible solution and the best is required, a particle represents an individual solution to the problem. The learning of the particles comes from two sources, one is from a particle's own experience called cognitive learning and the other source of learning is the combined learning of the entire swarm called social learning. The individual learning experience is represented as *(pBest)* and the combined learning experience is represented as *(gBest)* value. The *pBest* is the individual particle best experience while *gBest* value is the general best experience. It is the general experience that controls the behavior of the entire particle (Ghorpade-Aher and Metre, 2014). The general and individual experience is used to compute the speed to the new position. At any given time individual particle has these two basic things which are individual position $X(t)$ and velocity $V(t)$ and a memory which holds previous best result when applied to optimization problems, a typical PSO algorithm starts with the initialization of a number of parameters. One of the important initializations is selecting the initial swarm (Ghorpaede-Aher and Metre, 2014).

Together $pBest$ and $gBest$ are used to define the velocity of the particle which guides the particle towards a better solution. The velocity is calculated as given in equation 4.

$$V_i(t+1) = \omega x V_i(t) + c_1 r_1 (pBest_i(t) - X_i(t) + c_2 r_2 (gBest(t) - X_i(t) \quad (4)$$

Where $V_i(t)$ is the current velocity of the particles i while $V_i(t+1)$ is the new velocity that need to be achieved to be able to move from the current position to the new position. The range of velocities is bounded between V_{max} and V_{min}; Where V_{max} is the maximum velocity and $V_{min}$is the minimum velocity. The parameters c_1 and c_2 are the acceleration coefficients for social and cognitive components. The usual choice is to set C1 = C2 within the range [0,4]. $r_1$and r_2 are two random numbers ranging from 0 to1 that determines the influence of pBest and gBest on the velocity update formula, and ω is the inertia of the particle which controls the momentum of the particle. Velocity added to the current position provides the new position of the particle which is given by

$$X_1(t+1) = X_i(t) + V_i(t+1) \quad (5)$$

The value of the inertia weight w of the particle can be calculated using equation (6).

$$w = W_{max} * (\exp(-iter)) \quad (6)$$

Where W_{max} is the maximum value of w (0.9) and $iter$ is the current iteration number. In general, the inertia weight decreases linearly from 0.9 to 0.4 throughout the search process. The parameters r_1 and r_2 are modified based on the following equation

$$r(x) = \begin{cases} 0, & r(x) = 0 \\ frac(\frac{1}{x}) = \frac{1}{x} mod\ 1, & r(x) \epsilon\ (0,1) \end{cases} \quad (7)$$

Figure 1 shows how $pBest$ and $gBest$ affect the particles movement from position $X_i(t)$ to $X_i(t+1)$ while the PSO algorithm is shown in Figure 2 (Li Yeh *et al*, 2012).

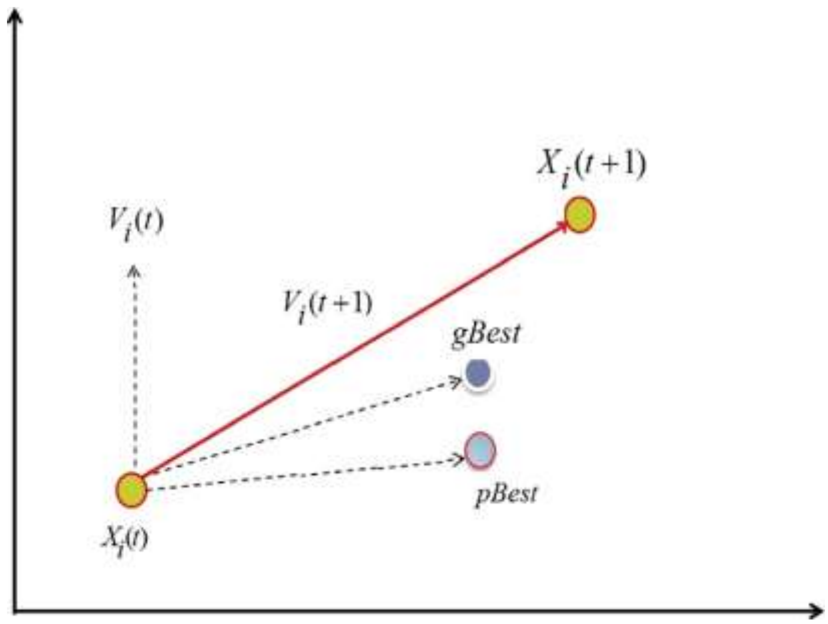

Fig. 1: Graphical representation of the particle repositioning (Source: Alam, *et al*, 2014)

The Algorithm of the PSO

Parameters: No of particles N, r1, r2, w, c1, c2, Vmax, Vmin

Initialize the velocity and position of the particle Vi(t) = 0 and Xi (t) = 0

Start

While within bounded area

Begin

For I = 1 to n

Evaluate fitness function and set pBest and gBest

Update the velocity and position of the particles

Next i

End While

End

In literature, many works have tried to optimize clustering algorithms using particle swarm optimization and other swarm intelligence techniques in an attempt to advance the efficiency and workings of these algorithms. First work that was reviewed was the one done by (Chuang, *et al*, 2012); in their work they use gauss chaotic map to prevent early termination of PSO at local optimum. In the application of their methodology the values gotten from the gauss-chaotic was used to replace r1 and r2. The cumulative of the intra cluster differences were used as the fitness function. They used six public UCI dataset to investigate the performance of the gauss PSO against other clustering algorithms namely; k-means, PSO, NMPSO, KPSO, KNMPSO. Evaluation results showed that their algorithm performed better than others in terms of error rate and convergence (Chuang,*et al*, 2012). Their work tries using PSO for clustering while this work used PSO to hybridized K-mode clustering algorithm to improve the performance.

Another of such work reviewed was the one done by (Behera,*et al*, 2012); they observed that k-means algorithm is a proficient algorithm for clustering; however, has a challenge of handling large dimensional data. They employed Principal Component Analysis (PCA) algorithm together with k-means algorithm to solve this issue of dataset dimensionality. They observed that the optimization of k-means alone gave much better result; and suggested the optimization of k-means with PSO technique together with enhanced PCA for grouping of dataset with high dimension (Behera,*et al*, 2012). Their work hybridized k-means with PCA while this work optimized K-mode with Particle Swarm Optimization.

Methodology for PSO Optimization of K-Mode Algorithm

The fitness value was evaluated using Equation

$$Fitness = \sum_{i=1}^{n} d(x_i, q_l) \tag{8}$$

Wherex_i is the ith element and q_i is the nearest cluster centre to x_i. K mode algorithm minimizes cost function, so the smaller the number the more similar the elements. The swarm size was set to equal the amount of data in the dataset (1733); 1233 instances were used to train the data while 500 instances were used to test the data. The PSO takes the fitness value as initial cost, the velocities and weights are picked at random and two centroids from each class were randomly picked. The loop runs updating the weights and velocities and for each data point the centroids are updated using the fitness value. When the centroids and weights stop updating, the final weights are used to cluster the test data. This work methodology is similar to the one in Zhao (2012) except in the nature of fitness function and nature of data set used. The algorithm for the PSOK-mode and flowchart are shown in Figure 3 and Figure 4 respectively.

Algorithm
Parameters: No of particles N, r1, r2, w, c1, c2, Vmax, Vmin *Initialize the velocity and weight of the particle Vi(t) = 0 and Xi (t) = 0* *Start* '$Fitness = \sum_{i=1}^{n} d(x_i, q_i)$ *While within bounded area* *Begin* *For I = 1 to n* *Cost= fitness,* *Set pBest and gBest* *Update the velocity and position of the particles* *Next i* *End While* *Kmode (test data,500)* *Evaluate Classification* *End*

Fig. 3: PSOKmode Algorithm

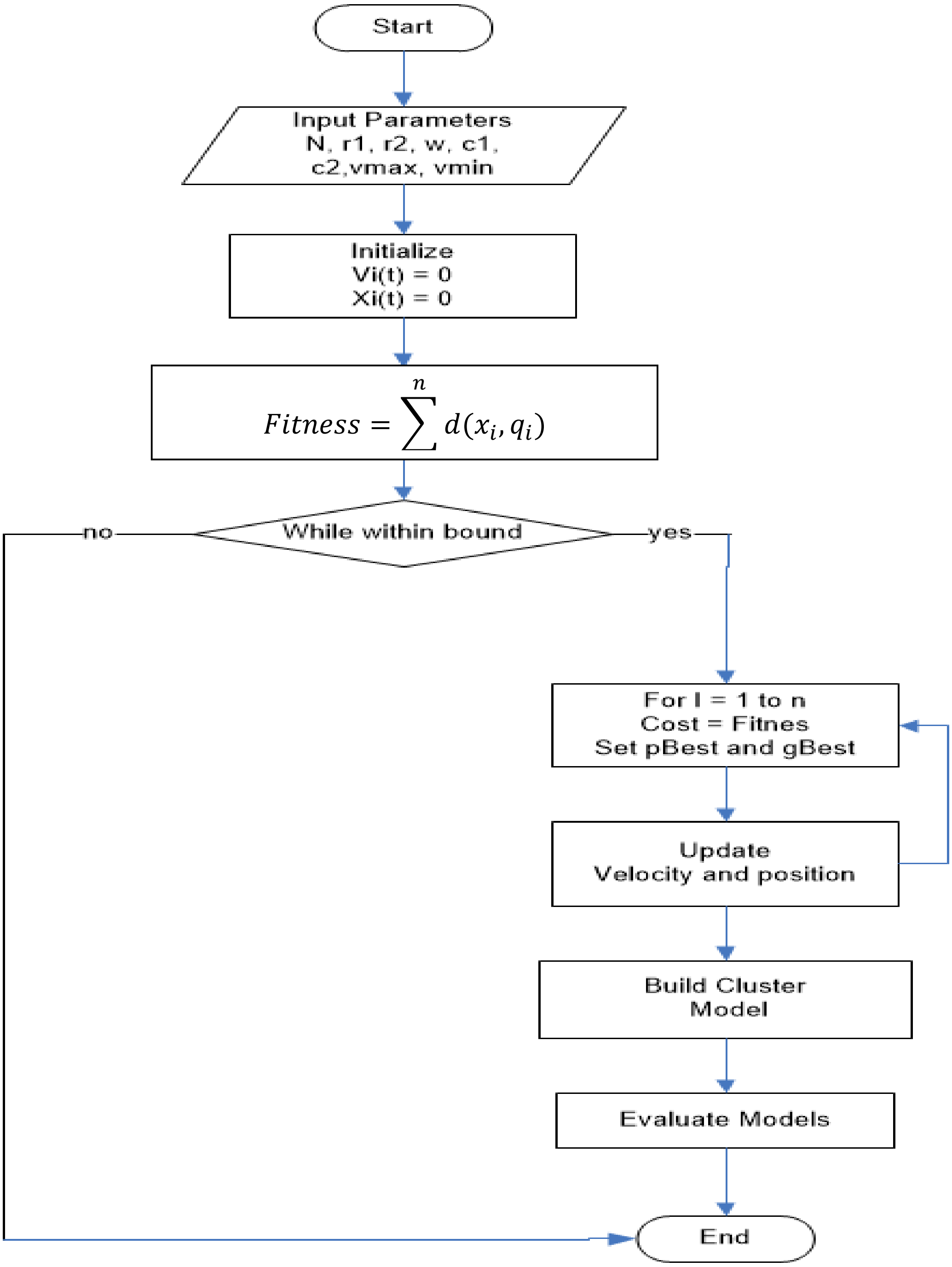

Fig. 4: PSOK-mode Flowchart

5.1. Table 1: Description of UCI and Crime Dataset used for viability test on PSOKM

Name	Records	Fields
Breast Cancer	699	10
CMC	1473	10
Wine	176	14
Iris	150	5
Glass	215	10
Crime dataset	1733	6

To test the viability of PSOKM five different datasets were chosen from UCI database as described in table 1.

Evaluation Metrics

The metrics used in the discussion of the result are as follows:

1) True Positive Rate (Sensitivity): this shows the statistics of correctly classified instances in each classification model

$$SN = \frac{TP}{TP + FN}$$

2) False Positive Rate (Specificity): is the report of instances incorrectly labeled as correct instances.

$$SP = \frac{TN}{TN + FP}$$

3) Accuracy: This shows the percentage of accuracy of the model

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}$$

4) Time: time taken to build the models
5) ROC curve: is used to visualize classifiers performance. It is usually plotted using two metrics: TP Rate and FP Rate. The y axis is usually for TP Rate while x axis denotes the FP Rate. The ROC area is used to measure its performance. If the area is 1 it indicates perfect prediction, if it is 0.5 it implies random guess

RESULT DISCUSSION

Comparison of Results Between K-mode and PSOK-mode on Crime data

Table 2: Tabulated Results of Viability of PSOKM using Crime dataset

Parameters	KM	PSOKM
Sensitivity	75.5	89.36
Specificity	83.33	33.33
Accuracy	76	89.36
Time	93.8 Secs	3.02 Secs

The tabulated result in Table 2 reveals that the PSOKM algorithm gave higher accuracy of 89.3 and lesser time of 3.02 secs compared to K-mode that gave accuracies of 76 and higher time of 93 secs respectively.

ROC Curve of Kmode and PSOKmode Algorithms

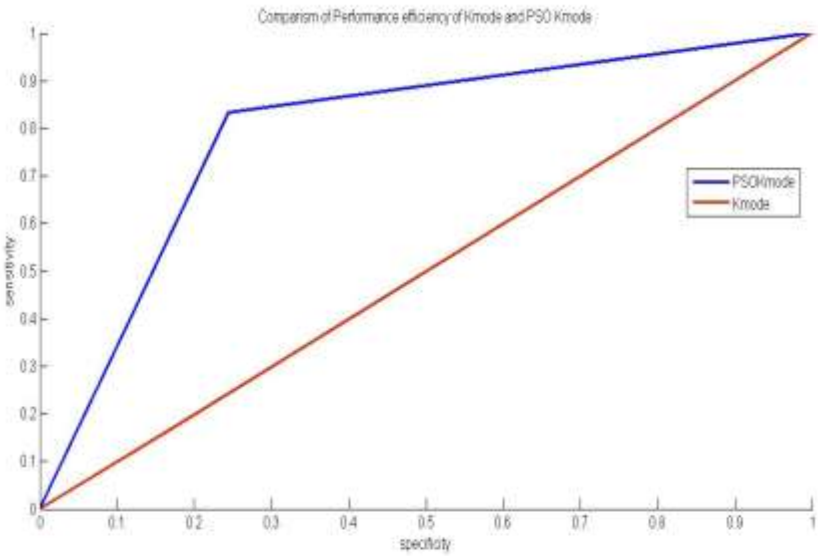

Figure 5: Comparison of Performance Efficiency of K-mode and PSOK-mode Algorithms Using ROC Curve.

In ROC curve graph, smaller the area under curve then lesser the efficiency of the algorithm. The ROC curve reveals that the PSOK-mode shows greater area under the curve which implies better performance.

Viability Test on PSOKM against other Variants Algorithms Using UCI dataset

The viability study was in comparison with these algorithms: Hierarchical PSO clustering (HPSO), Hierarchical Agglomerative cluster (HAC), K-mean Harmonic Means (KHM), PSO K-Harmonic Means, Hybrid PSO (Hybrid), K-means PSO based Nelder-Mead (NM) simplex method (K-NM-PSO). The results are as shown in Figure 6, Figure 7 and Figure 8 respectively

Viability Test Results of PSOKM on Accuracy

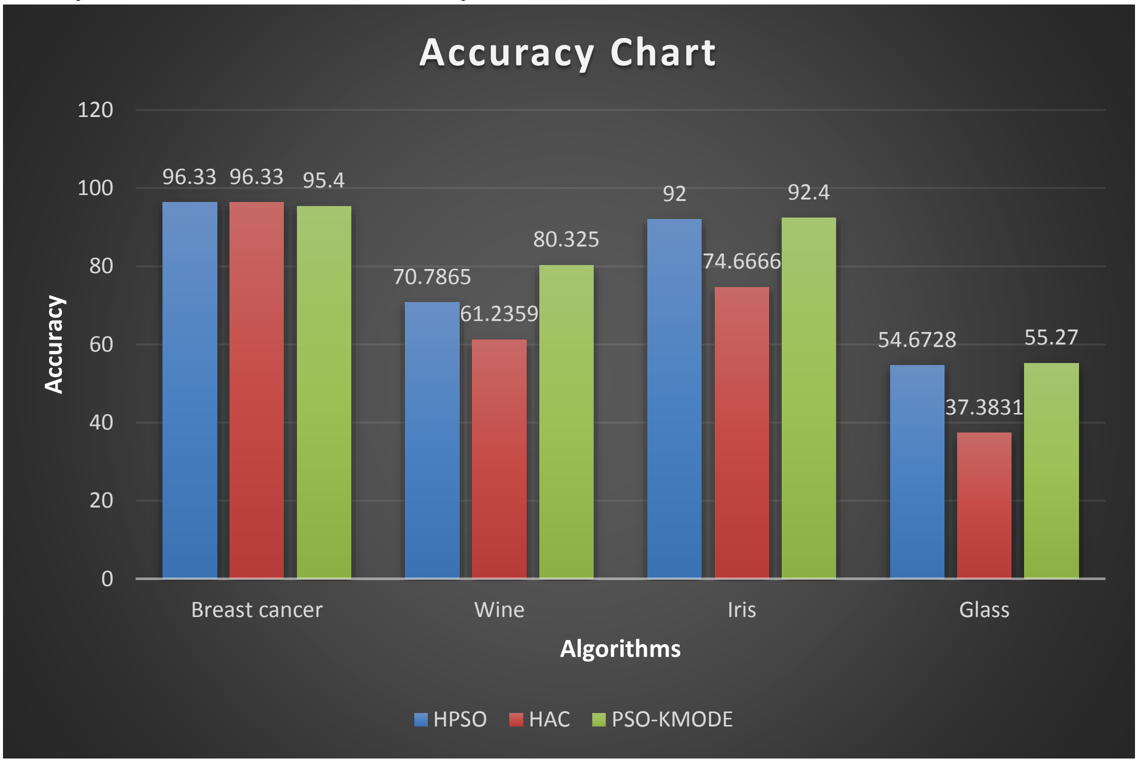

Figure 6: Viability Test Results of PSOKM on Accuracy

This accuracy graph reveals that PSOKM has higher accuracy compared to other algorithms on most of the dataset except on breast cancer data.

Viability Test on Run Time for PSOKM

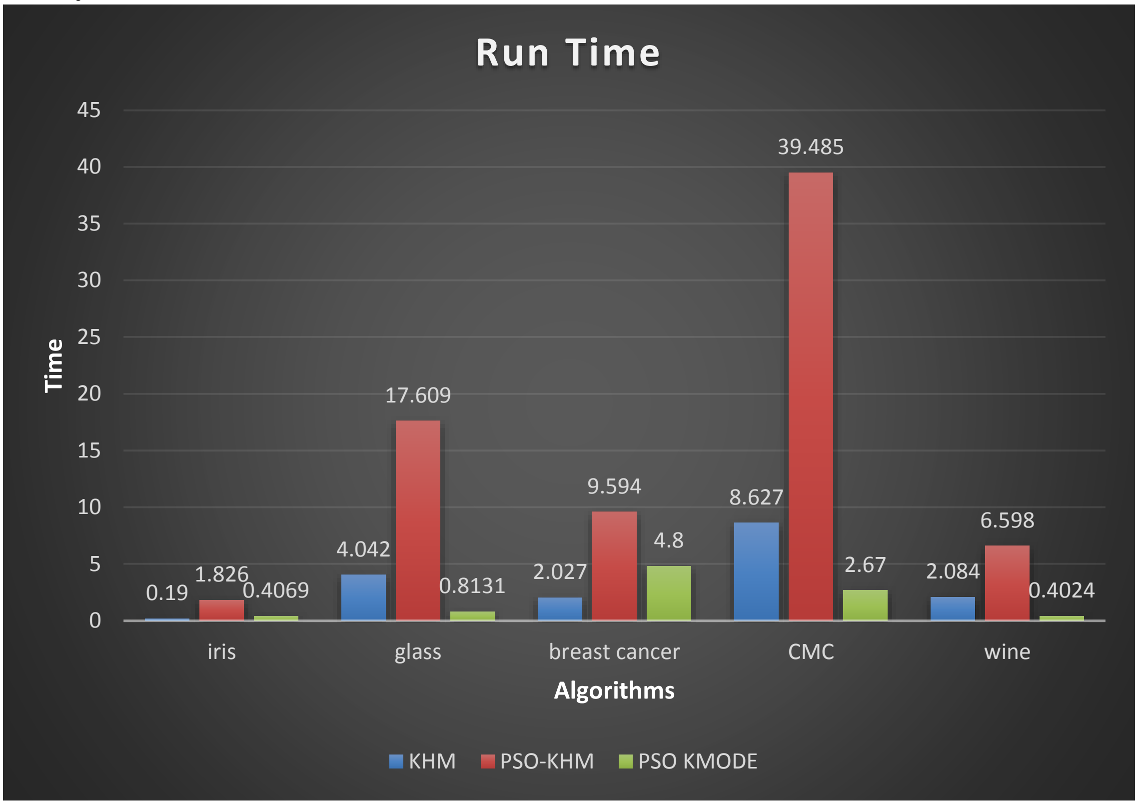

Figure 7: Viability Test on Run Time for PSOKM

The results revealed that PSOKM algorithm takes lesser time in building its model compared to other algorithms on most dataset.

Inter Cluster DistanceTest on PSOKM

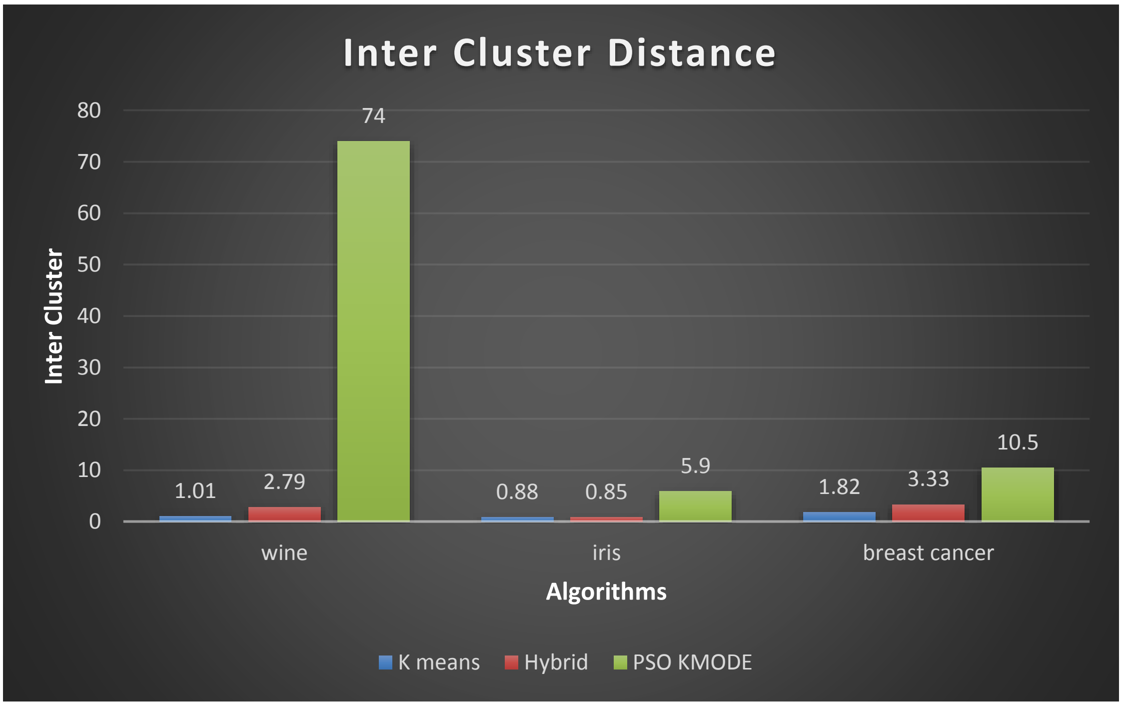

Figure 8: Inter Cluster Distance Comparison

The comparative graph in Figure 8 showed pictorially that the inter cluster distance of PSOKM is considerably greater than that of other algorithms compared with.

Intra Cluster DistanceTest on PSOKM

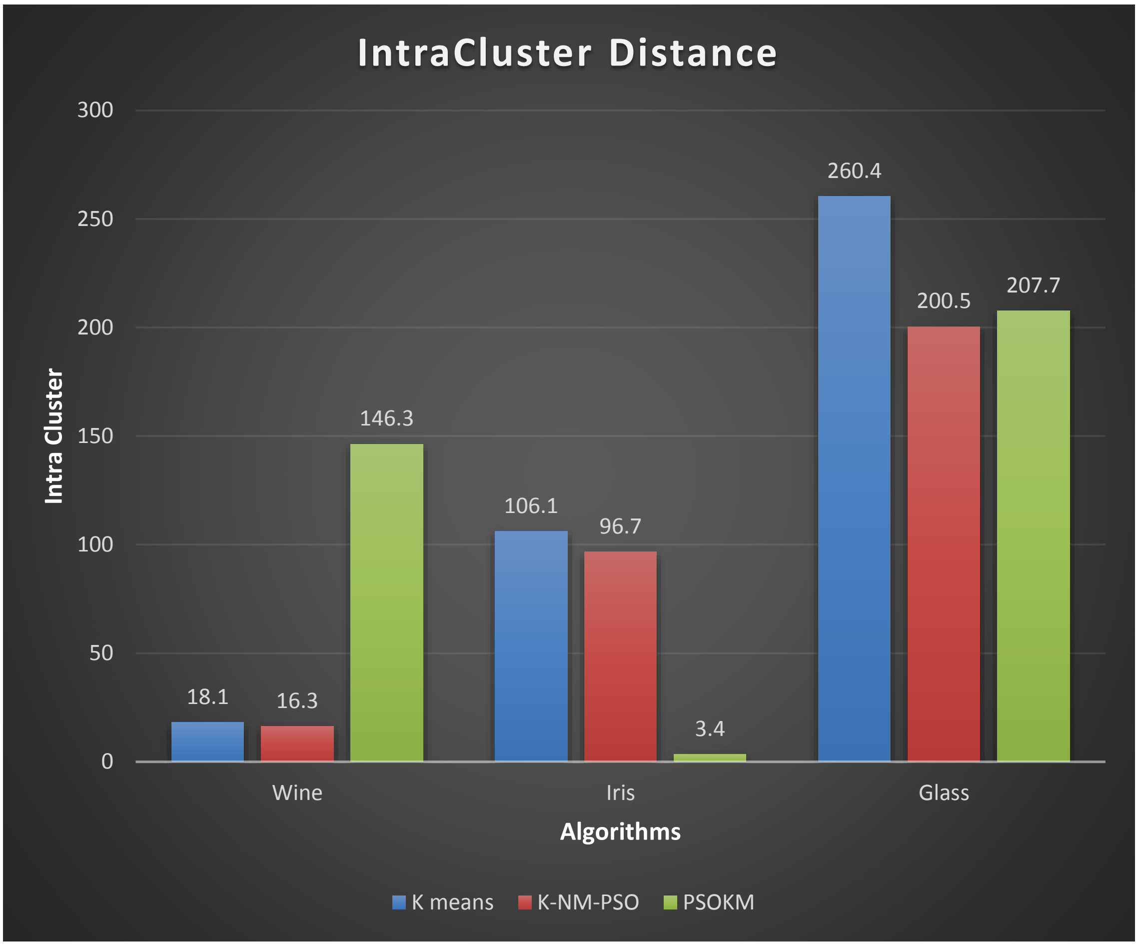

Figure 9: Intra Cluster Distance Comparison

The comparative graph in Figure 9 showed pictorially that the intra cluster distance of PSOKM is considerably smaller than that of other algorithms except in wine dataset.

CONCLUSION

Particle Swarm Optimization (PSO) is one of the Swarm Intelligence techniques whose algorithm has been proved to improve efficiency and accuracy in many area of human endeavour where it has been applied. This research applied the PSO technique to improve the efficiency and accuracy of K-mode clustering algorithm. The evaluation result revealed that optimizing the algorithm improved the accuracy from 76 % to 89.3%. The ROC curve values which is used to measure performance of classification showed higher area under curve in optimized K-mode (PSOKM) than in ordinary K-mode.

The optimized algorithms when used for classification prove to have reduced the classification time from 93.8 sec to to 3.02 secs . This work used five UCI benchmark dataset to conduct viability test for PSOKM. These datasets are frequently used datasets in data mining and machine learning approaches that have shown reasonable success in their applications. The results of the viability tests carried out on the proposed method PSOKM demonstrated that the accuracy of this method (PSOKM) is much better in most dataset. The method also outperformed the other variants algorithms in most of the evaluation metrics under consideration. Most works used in this comparison are coded in MATLAB only, few are coded in FORTRAN thus the complexity of the methods are considered to be the same.

REFERENCES

Alams, S., Dobbie, G., Koh, Y. S., Riddle, P., & Ur Rehman, S. (2014). Research on particle swarm optimization based clustering: a systematic review of literature and techniques. *Swarm and Evolution Computation*, 17, 1-13.

Behera, H. S., Abhishek, G., & Sipak, K. M. (2012). A new Improved Hybridized K-means Clustering Algorithm with Improved PCA Optimized with PSO for High Dimensional Dataset. *International Journal of Soft Computing and Engineering*, 2(2), 2231-2307.

Chuang, L. Y., Lin, Y. D., & Yang, C. H. (2012). An improved particle swarm optimization for data clustering. In *Proceedings of the International Multi Conference of Engineers & Computer Scientist,* 1(1).

Ghorpade-Aher, J., & Metre, V. A. (2014). Clustering Multidimensional Data with PSO based Algorithm. *arXiv preprint arXiv:1402.6428*.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data setsin data mining. *In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 1-8.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets withcategorical values. *Data Mining and Knowledge Discovery*, 2(3): 283-304.

Huang, Z., & Ng, M. K. (2003). A note on k-modes clustering. *Journal of Classification*, *20*(2), 257-261.

Jiawei, H., Micheline, K., & Jian, P. (2012). *Data mining: Concept and Techniques* (3rd ed).San Francisco, Elsevier.

Li-Ye, C., Yu,-D. L., & Cheng-Hong, Y. (2012). An improved PSO for data clustering. *Proceedings of the International Conference of Engineers and Computer Scientist*, 1, 26-40.

Martens, D., Baesens, B., & Fawcett, T. (2011). Editorial survey: swarm intelligence for data mining. *Machine Learning*, *82*(1), 1-42.

Satyobroto, T. (2011). Mathematical modelling and applications of particle swarm optimization. *Doctoral Dissertation, Institute of Technology, Blekinge*.

Zhang, Y., Fu, A. W., Cai, C. H., Heng, P. A. (2000). Clustering categorical data. *In: Proc of ICDE'00*, pp. 305-305, 2000

Zhao, X., & Mei, L. (2013). 3D object retrieval based on PSO-K-mode. *Journal of Software, 8(4), 963-970.*