International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:2, 2010

Improving RBF Networks Classification Performance by using K-Harmonic Means

Z. Zainuddin, and W. K. Lye

***Abstract*—**In this paper, a clustering algorithm named K-Harmonic means (KHM) was employed in the training of Radial Basis Function Networks (RBFNs). KHM organized the data in clusters and determined the centres of the basis function. The popular clustering algorithms, namely K-means (KM) and Fuzzy c-means (FCM), are highly dependent on the initial identification of elements that represent the cluster well. In KHM, the problem can be avoided. This leads to improvement in the classification performance when compared to other clustering algorithms. A comparison of the classification accuracy was performed between KM, FCM and KHM. The classification performance is based on the benchmark data sets: Iris Plant, Diabetes and Breast Cancer. RBFN training with the KHM algorithm shows better accuracy in classification problem.

***Keywords*—**Neural networks, Radial basis functions, Clustering method, K-harmonic means.

I. Introduction

RADIAL Basis Function Networks (RBFNs) is a class of neural network which has attracted a lot of interest of researchers due to its simple structure, well established theoretical basis and fast learning speed. RBFNs are applicable to different fields such as function approximation, regulation, noisy interpolation, density estimation, optimal classification theory and potential functions [1]. RBFNs form a unifying link between the fields above and this cause the training in RBFNs substantially faster than the methods used to train Multilayer Perceptron networks (MLPNs). Hence, RBFNs represent an alternative to the widely used MLPNs.

The training in RBFNs can be classified into two stages: (i) the basis function parameters (corresponding to hidden units). Typically, fast and unsupervised clustering methods are used to determine these parameters (ii) the weights in final layer. A linear system solution involves in this weight determination.

To design an ideal architecture of the network is an issue in the neural network community. One of the advantages of RBFNs compared to MLPNs is possibility of choosing suitable parameters for the hidden units without involving non-linear optimization of the network parameters. However, the performance of RBFNs depends critically on the number, the position and the shapes of the hidden units [2]. The general way to select the centre of hidden units is to superpose each centre to the data set point. This method needs a heavy computational cost and leads the network to poor generalization [3].

There have been several existing strategies proposed to select the centre of hidden units. Among them are random selection, systematic seeding, expert contribution, clustering, editing methods. A clustering method is one of the most common methods. K-means (KM) and fuzzy c-means (FCM) are two popular centre based algorithms that have been developed to solve the clustering problem. Both algorithms are used to determine the centre of hidden units. In this paper, we propose K-harmonic means which is another alternative clustering method in the determination of RBFN centers.

This paper is organized as follows: Section II reviews RBFN and their training algorithm while section III focuses on the clustering methods details. A discussion on the experimental result is given in section IV while conclusions are presented in section V.

II. Radial Basis Function Network (RBFN)

A RBFN is a three layer feed forward neural network. The first layer (input layer) consists of n input units, **x**. The second layer (hidden layer) introduces a set of basis functions, one for each input unit and sets the weights for the linear combination of basis functions. The basis functions are non-linear functions, $\xi(\boldsymbol{x})=(\|\boldsymbol{x}-\boldsymbol{x}_i\|)$, of the input vector $\mathbf{x}_i$. The linear function can be written as:

$$\sum_{i=1}^{n} \boldsymbol{w}_i \xi(\|\boldsymbol{x}-\boldsymbol{x}_i\|) \quad (1)$$

The third layer (output layer) provides outputs, $y=\varphi(\xi)$, which is a simply a weighted linear summation of the outputs from the second layer as shown in equation (1). A simple RBFNs structure is provided in Fig. 1.

The RBFN is activated by, in common cases, the Gaussian basis function:

$$\xi(\boldsymbol{x})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}\|^2}{2b^2}\right) \quad (2)$$

where **x** is the input vector, **c** is the centre of basis function and b is the width. Each basis function gives its higher output when the input is close to the centre and the value decreases monotonically as the distance from the centre increases. There are a variety of measurements to evaluate the distance, but the Euclidean distance is the most popular one and is also used here.

Z. Zainuddin is with Universiti Sains Malaysia, 11800 Penang, Malaysia (phone: +604-6532510; fax: +605-6570910; e-mail: zarita@cs.usm.my).

W. K. Lye is with Universiti Sains Malaysia, 11800 Penang, Malaysia (e-mail: kitlye@yahoo.com).

As mentioned above, the RBFN is accomplished into two stages: (i) training of the centers in the hidden layer which is influences the performance of RBFN, (ii) calculation of the hidden to output weights by solving a linear system. The linear system can be solved to yield

$$\boldsymbol{w} = \xi^{-1}\boldsymbol{y} \tag{3}$$

where $\boldsymbol{w}$ is the output weights. Let d be the target output of the input x, where the network is designed in $\Re^n \rightarrow \Re^m$. m is the size of the output. This is a training set of the network. In every training procedure, the network computes the actual output, φ and the error e of each output unit by $e = d - \varphi$. The goal of the RBFN learning is to minimize the error function, called training error. This leads us to an optimization problem:

$$E = \frac{1}{2}\sum_{n}\left(d - \varphi\right)^2 \tag{4}$$

An important measure of the trained network performance is the generalization error computed over a set of data which is never involved in the training, so called testing data.

III. Clustering Method

In this section, we recall the K-means (KM) and Fuzzy c-means (FCM) clustering algorithms which are the popular algorithms used in the determination of centre in the hidden units. Moreover, K-harmonic means is introduced at the end of this section.

A. K-means Clustering Methods

K-means is the one of the simplest clustering method [4]. It is an algorithm to find K-centers of the data set based on the dissimilarity (distance) of the data set to the centers. It takes K numbers of initial value as a starting point to find the cluster centers. In the algorithm, the problem is defined to minimize the distance between data and the corresponding cluster centres [4]. Normally, the total means square quantization error (MSE) is used to measure the performance of KM clustering method. Besides simplicity, KM is also well known in its speed in clustering of a large data set. However, KM faces two drawbacks: (i) performance highly depends on initial state (initial centre) and (ii) convergence to local minima. Different initial state provides different clustering result. This is because the algorithm often converges to local minima. The problem can be seen obviously when the initial centres are not well separated.

B. Fuzzy c-means

Fuzzy c-means was originally developed by Dunn in 1973 [5] and generalized by Bezdek in 1974 [6]. The idea of FCM is to smooth the hard nature of the KM algorithm which a data only assign to a cluster. For example, the KM algorithm with three clusters only allow the data to assign to one of the three clusters. In contrast, FCM algorithm employs the fuzzy partitioning to let the data can belong to all clusters with different membership grade between 0 and 1 and the sum of the membership grade is 1. With the highest membership grade the data is assign to the corresponding cluster [7].

Unfortunately, the same drawbacks happen to FCM algorithm. FCM algorithm cannot ensure that it converges to global optima. The clustering result also depends on the initial membership grade because the cluster centers are initialized using membership grades which are randomly initialized.

C. K-harmonic Means

K-harmonic means (KHM), like KM and FCM algorithm, is a centre based clustering algorithm. It is proposed by Zhang [8,9] and modified by Hammerly and Elken [10]. The harmonic means is defined as

$$HM\left(\{a_1 \ldots a_k\}\right) = \frac{K}{\sum_{k=1}^{K}\frac{1}{a_k}} \tag{5}$$

Equation (5) has a characteristic that if one of the values is small, the output of HM will be small. Conversely, if none of the values of *a* are small then the HM will be large. We assign the data to the center by considering the minimum distance in KM and HM. Hence, HM can be used as a minimum function,

$$HM\left\{\left\|\boldsymbol{x} - \boldsymbol{c}\right\|^2 \middle| \boldsymbol{c} \in \boldsymbol{C}\right\} = \frac{\left|\boldsymbol{C}\right|}{\sum_{c \in C}\frac{1}{\left\|\boldsymbol{x} - \boldsymbol{c}\right\|^2}} \tag{6}$$

where c is the cluster centre and x is the data. We can use (6) to calculate the distance between the data and the cluster centres. Equation (6) can be incorporated into the performance function as:

$$\Phi\left(\left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|^2\right) = \sum_{i=1}^{n}\frac{k}{\sum_{i=1}^{k}\frac{1}{\left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|^p}}, \text{ j = 1,…,k} \tag{7}$$

where k is the number of clusters. The value of p is associated to the power of distance calculation. According to the KM algorithm, p should be equal to 2 while the distance calculation is based on squares distance. However, previous work has shown that p>2 works better in KHM [9].

As mentioned above, KM algorithm assigns a hard membership to data which let the data belong to exactly one cluster centre. This will cause the data only has its influence to a particular cluster centre. Moreover, in the high local density area of data points and centers, the centre maybe unable to move away from the area despite a centre is needed nearby. Too many centers crowd at certain area may cause the worse local solution. Hence, reposition the centre may give better global effect and beneficial to the clustering. However, KM algorithm cannot perform the centre swapping.

FCM and KHM are designed to use membership grade to overcome the swapping problem. The membership grade in KHM can be defined as:

$$m\left(\frac{\boldsymbol{c}_j}{\boldsymbol{x}_i}\right) = \frac{\left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|^{-p-2}}{\sum_{j=1}^{k}\left\|\boldsymbol{x}_i - \boldsymbol{c}_j\right\|^{-p-2}} \tag{8}$$

KHM algorithm is sensitive to the fact that there exist two or three centers to a data point. The algorithm will remove one or more these centers to area where the data points do not have

close centre. The objective function uses the distance of data to all centers. With the swapping property, KHM will lower the value of the objective function. In KM and FCM, the objective function gives equal weight to all of the data points. KHM assigns different weights to different data points based on the following weight function:

$$w(\boldsymbol{x}_i) = \frac{\sum_{j=1}^{k} \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|^{-p-2}}{\left(\sum_{j=1}^{k} \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|^{-p} \right)^2} \quad (9)$$

From (9), we understand that the harmonic mean will assign a large weight to a data point that is not close to the any centre and a small weight to a data point that is close to one or more centers. By increasing the weight of data which is not close to any centre, the algorithm can attract centers in a dense area without changing the weight of the data points of the area. This can solve the problem where cluster centres crowd at dense areas. This property makes the KHM algorithm less sensitive to the initial cluster centre which assign randomly. The cluster centres are updated as follow:

$$\boldsymbol{c}_k = \frac{\sum_{i=1}^{n} \frac{1}{d_{i,k}^{p+2} \left(\sum_{j=1}^{k} \frac{1}{d_{i,j}^{p}} \right)^2} \boldsymbol{x}_i}{\sum_{i=1}^{n} \frac{1}{d_{i,k}^{p+2} \left(\sum_{j=1}^{k} \frac{1}{d_{i,j}^{p}} \right)^2}} \quad (10)$$

IV. Experimental Results and Discussion

All the following experiments are performed using Matlab® with a neural network toolbox. All computing were performed on a computer Intel Core 2, 1.83 GHz, 1GB RAM. The proposed algorithm is performed on five benchmark data sets. The five bench mark data sets are Iris Plant, Diabetes, Breast Cancer, Hepatitis and Lung Cancer data set. All the data sets are identified as classification data sets. The comparison was made on the accuracy of classification in the RBFN.

In the RBFN architecture, KM, FCM and KHM are applied in all the cases to determine the prototypes which are then used to initialize the basis function centre. In order to simplify the calculation, we set the centres width (spread value) to be 1 in all the cases. Different values of p in the KHM algorithm lead to different performance. We run the empirical trials and concluded that the best value for p is in the range of 2 to 3.5. In this experiment, we used value of p as 2.8 in all cases.

The performance of RBFN may depend on the selection of the training and testing sets. There was a little variation in the results when different training and testing sets were used in the RBFN. This implies that the convergence of the optimum parameters was obtained in each case. To solve this problem we try to apply k-fold cross validation in the data set. It means that we let the data set separated to k disjoint subsets. In the method, k-1 folds were used for training and the last fold was used for testing. This process was repeated k times, leaving a different fold for testing each time. 100 independent trials were applied to each data set. The average performance was considered in the experiments.

The first data set is Iris Plant data set. It has 150 samples with four variables in each sample. The data set represents three classes of iris plants with 50 samples each. The second data set is Diabetes which consists of 768 samples with eight variables in each sample. There are two classes in the data set. The third data set is the Breast Cancer data set which consists of 569 samples with 30 variables in each sample. The data set represents two classes of diagnostic field – benign and malignant. The forth dataset is Hepatitis data set which consists of 155 samples with 19 variables in each sample. There are two classes of survival outcome – die or live. The fifth data set is Lung Cancer data set which consists of 32 samples with 56 variables in each sample. There have three types of pathological lung cancer as the outcome in this data set. [11].

The differences of the experimental results from the three clustering based RBFN are examined using statistical method – analysis of variance (ANOVA). We set the statistical significant level to 0.05 and posthoc test was done on all possible pairs of group accuracy means. In this case, there are KHM-RBFN versus KM-RBFN, KHM-RBFN versus FCM-RBFN and KM-RBFN versus FCM-RBFN. The results have shown that statistically significant in the difference between the accuracy mean of KHM-RBFN versus KM-RBFN and KHM-RBFN versus FCM-RBFN. Most of the p-values are <0.001. However, the KM-RBFN versus FCM-RBFN has shown statistically not significant in comparison of accuracy mean. The statistical analysis was performed using PASW Statistics (formerly known as SPSS) version 17.0.2.

The different structure of RBFNs is designed by varying the number of hidden nodes, i.e. sample size-N-number of classes, N = {2, 3, … , 8}. For particular, Iris Plants has 4 - 2 - 3 structure with two hidden nodes.

Table II shows the accuracy of classification for the Iris Plants data set. Among the three clustering algorithms, the KHM-RBFN gave the highest accuracy and lowest standard deviation for different number of clusters.

Table III shows the accuracy of classification for the Diabetes data set. KHM-RBFN gives the highest accuracy for different number of clusters but not the lowest standard deviation. The RBFN show the consistent accuracy in two clusters model in each clustering algorithm. Table 4 shows the accuracy of classification for Breast Cancer data set. The KHM-RBFN attained the best accuracy among the three clustering algorithms.

Table IV shows the accuracy of classification for Hepatitis data set. As expected, KHM-RBFN gives the highest accuracy among the three models. To highlight here is the relatively low standard deviation in KHM-RBFN compare to the other two models. Table 5 shows the accuracy of classification for Lung Cancer data set. Despite the RBFN gives the low accuracy, KHM-RBFN still the highest among the three.

Figs. 1-5 are the plot of accuracy mean versus number of cluster. From the plots we can see the KHM-RBFN is more superior to the KM-RBFN and FCM-RBFN in term of accuracy. Moreover, we also can notice that accuracy of KHM-RBFN shows an uptrend in Iris, Diabetes and Breast Cancer data set. It may due to the more appropriate clusters are selected in the KHM-RBFN compare to another two models

V. Conclusion

A clustering algorithm, namely k-harmonic means, was implemented into RBFN to search the centroids of hidden units of RBFN. K-harmonic means is a clustering algorithm which is less sensitive to the initial centres compare to conventional clustering algorithms: fuzzy c-means and k-means. Five benchmark data sets were used in the classification. The RBFN implemented with k-harmonic means shows improved performance in the classification problems with respect to fuzzy c-means, k-means and standard RBFN.

References

[1] C.M. Bishop, Neural Networks for Pattern Recognition. Oxford University Pres, New York, USA, 1995.
[2] N. Benoudjit, M. Verleysen, On the kernel width in radial basis function networks. Neural Processing Letters 18, 2003, 139-154.
[3] J. Moody, C.J. Darken, Fast learning in networks of locally-tuned processing unit. Neural Computation 1, 1989, 281-294.
[4] K.Warwick, J.D. Mason and E.L. Sutanto, Neural network basis function center selection using cluster analysis. Proceeding of American Central Conference, Washington, June, 1995
[5] J.C. Dunn, A fuzzy relative of the ISODATA process and its use n detecting compact well-separated clusters. J. Cybernet. 3, 1973, 32-57.
[6] J. Bezdek, Pattern recognition with fuzzy objective function algorithm, Plenum Pres, NewYork, 1981.
[7] R.L. Canon, J.Dave and J.C. Bezdek, Efficient implementation of the fuzzy cmeans clustering algorithms. IEEE Trans Pattern Arial Machine, Intell 8, 248-255.
[8] B. Zhang, M. Hsu, U. Dayal, K-harmonic means – a data clustering algorithm, Technical Report HPL-1999-124, Hewlett –Packard Laboratories, 1999.
[9] B. Zhang, M. Hsu, U. Dayal, K-harmonic means, in: International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM2000, Lyon, France, 12 September 2000.
[10] G. Hammerly, C. Elken, Alternatives to the K-means algorithm that find better clusterins, in: Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, pp. 600-607.
[11] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, 2008, http://archive.ics.uci.edu/ml/databases.html.

TABLE I
Accuracy Mean (Standard Deviation) of Classification for Iris (%)

Clustering algorithms	Number of clusters						
	2	3	4	5	6	7	8
KHM	94.70(0.90)	95.03(0.75)	95.03(0.78)	95.07(0.77)	95.09(0.76)	95.12(0.72)	95.15(0.79)
KM	93.57(1.31)	94.05(0.87)	94.08(1.02)	94.09(0.96)	94.10(1.18)	93.97(1.12)	93.99(1.06)
FCM	93.63(1.17)	94.11(0.82)	94.12(0.89)	94.08(0.91)	94.06(1.01)	94.06(1.03)	94.07(1.01)

TABLE II
Accuracy Mean (Standard Deviation) of Classification for Diabetes (%)

Clustering algorithms	Number of clusters						
	2	3	4	5	6	7	8
KHM	65.10(0.00)	63.58(0.39)	65.15(0.65)	65.77(0.84)	65.60(0.75)	65.55(0.94)	65.91(0.81)
KM	65.10(0.00)	63.43(0.38)	63.56(0.38)	63.53(0.41)	63.56(0.35)	63.54(0.42)	63.56(0.46)
FCM	65.10(0.00)	63.35(0.44)	63.52(0.40)	63.57(0.42)	63.56(0.36)	63.52(0.45)	63.58(0.36)

TABLE III
Accuracy Mean (Standard Deviation) of Classification for Breast Cancer (%)

Clustering algorithms	Number of clusters						
	2	3	4	5	6	7	8
KHM	88.30(0.28)	91.68(0.45)	92.09(0.17)	92.26(0.27)	92.15(0.21)	92.10(0.22)	92.36(0.31)
KM	88.47(0.30)	91.12(0.59)	90.76(0.15)	90.98(0.13)	90.97(0.10)	90.97(0.12)	90.94(0.15)
FCM	88.44(0.30)	89.28(0.64)	91.27(0.18)	91.28(0.17)	91.26(0.19)	91.24(0.17)	91.28(0.16)

TABLE IV
Accuracy Mean (Standard Deviation) of Classification for Hepatitis (%)

Clustering algorithms	Number of clusters						
	2	3	4	5	6	7	8
KHM	79.34(0.14)	79.35(0.07)	79.32(0.22)	79.33(0.13)	79.30(0.25)	79.28(0.22)	79.31(0.27)
KM	78.63(0.21)	78.37(0.81)	78.40(0.53)	78.51(0.41)	78.42(0.51)	78.36(0.57)	78.12(0.72)
FCM	78.55(0.36)	78.52(0.37)	77.96(1.00)	78.40(0.52)	78.43(0.47)	78.24(0.61)	78.07(0.87)

TABLE V
ACCURACY MEAN (STANDARD DEVIATION) OF CLASSIFICATION FOR LUNG CANCER (%)

Clustering algorithms	Number of clusters						
	2	3	4	5	6	7	8
KHM	48.44(7.80)	48.25(7.44)	47.50(7.31)	47.38(7.95)	49.03(6.73)	46.35(8.71)	48.10(7.31)
KM	44.38(8.09)	43.28(6.92)	44.28(8.28)	41.44(6.80)	43.00(7.73)	41.35(7.05)	42.03(8.01)
FCM	44.50(5.89)	43.78(6.66)	43.85(6.49)	40.82(5.49)	43.22(6.30)	42.35(5.03)	45.00(5.89)

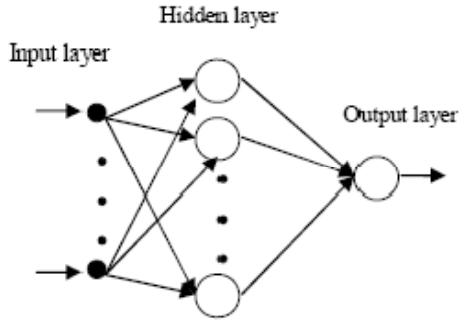

Fig. 1 A simple RBFN structure

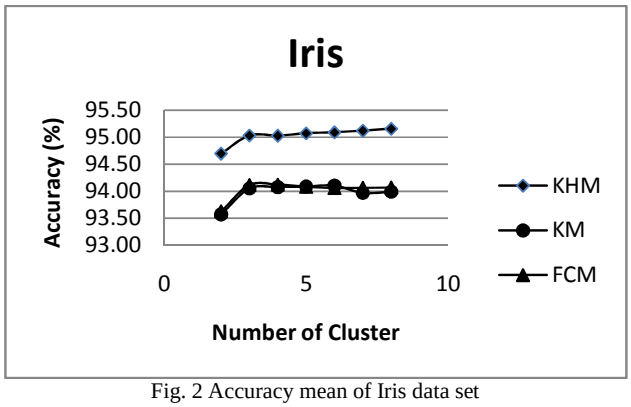

Fig. 2 Accuracy mean of Iris data set

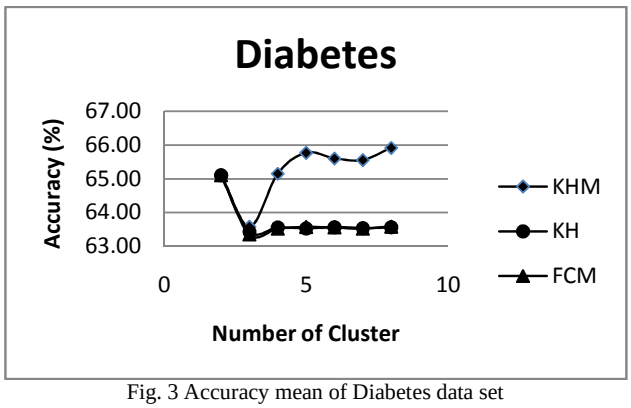

Fig. 3 Accuracy mean of Diabetes data set

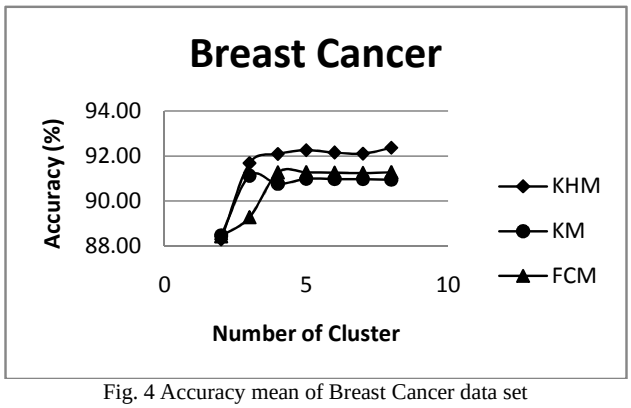

Fig. 4 Accuracy mean of Breast Cancer data set

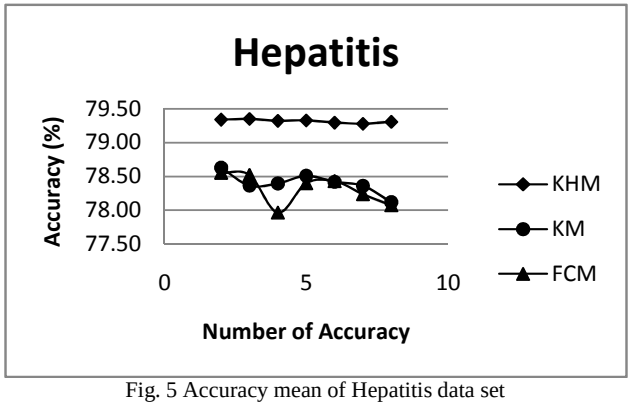

Fig. 5 Accuracy mean of Hepatitis data set

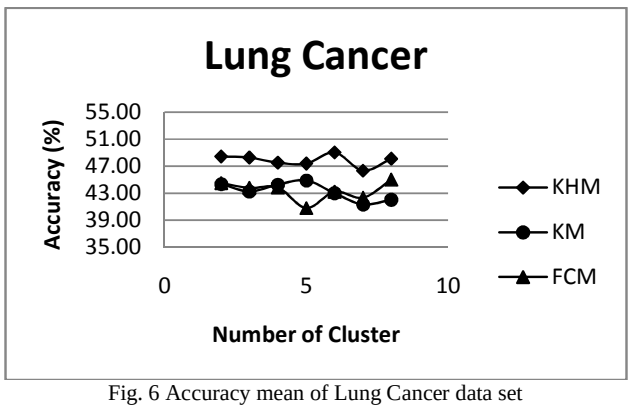

Fig. 6 Accuracy mean of Lung Cancer data set