Article can be accessed online at http://www.publishingindia.com

A MODIFIED K-MEANS ALGORITHM THAT DETERMINES NUMBER OF CLUSTERS AUTOMATICALLY

Jyoti Sharma*, G. N. Purohit**

**Banasthali University, Banasthali, Rajasthan, India. Email: jsharma.f41@gmail.com*
***Dean, AIM & ACT, Banasthali University, Banasthali, Rajasthan, India. Email: gn_purohitjaipur@gmail.com*

Abstract *This paper presents a modified k-means that not only increases classification accuracy but also determines optimal number of clusters automatically. It is a two phase clustering algorithm based on k-means, membership degree and standard deviation. Phase I automatically determines the number of clusters. Standard deviation of membership value helps us to identify number of clusters automatically. The phase II increases the classification accuracy. It does not require number of clusters as parameters like other previous most widely used unsupervised pattern clustering algorithms such as Fuzzy C-means and K-means. Experiments were done on some benchmark datasets and synthetic datasets to evaluate the performance of proposed approach.*

Keywords: *Clustering, Validitymeasures, K-means, Membership Degree*

Introduction

Classification of datasets is an important research area in the field of pattern recognition. K-means [1] and Fuzzy C-means [2] are the most widely used fuzzy clustering algorithms in pattern recognition applications. Both of them are good clustering algorithms but they must know the actual number of clusters in advance.

There are several modified versions of these algorithms were proposed in the literature such as FBSA-FCM based splitting algorithm [3], A-FCM-adaptive fuzzy clustering [4], PKM [5], simple & fast GKM algorithm [7], modified global k-means for minimum sum-of-squares clustering problems [8], FRBC-Fuzzy Rule Based Clustering Algorithm [9] etc.

Model selection algorithm is the algorithm that executes the existing algorithm multiple times and chooses appropriate result from them on the basis of validity criteria. These algorithms can be used as model selection algorithm to get the better partition result [3-6]. But this process is time consuming. There are several model selection methods available in the literature such as Akaike Information Criterion (AIC) [29], Bayesian information criterion (BIC) [29] [30] [31], Guaranteed Risk Minimization (GRM) [32], Minimum Description Length Principle (MDL) [32] and cross-validation [32] [33].

Different validity indexes were used to solve the above-mentioned common clustering problem. These validity indices were based on different factors such as inter-cluster and intra-cluster distance, sum of cluster variance [4], inter-cluster variation and intra-cluster [10], robustness based on mean and median [11] etc. There are several validity indexes or criteria have been proposed in the literature. Some of them are the Partition Coefficient [12], Partition Entropy [13] [14], Modified Partition Coefficient [15], Krzanowski & Lai [16], Silhouette [17], Calinski-Harabasz [18], Davies-Bouldin [19], Non Fuzzy Index [20] etc.

Two-phase clustering algorithm was proposed in this paper. The first phase is based on k-means and proposed validity index based on membership degree and standard deviation. This phase automatically identifies the number of clusters and also gives good clustering results and the second phase increases clustering accuracy.

This work consider k-means due to its simplicity and effectiveness for classification of data sets and fuzzy c-means's membership function to calculate similarity value or membership degree and standard deviation to identify the compactness of the clusters.

In order to improve the performance of k-means proposed work used membership function to calculate similarity and standard deviation of similarity value to identify number of clusters automatically.

The combination of all these solves the problem to provide number of clusters initially as parameter.

The reminder of the paper is organized as follows: Section 2 presents state-of-art. Section 3 focuses on the detailed description of proposed methodology. Section 4 depicts experimental analysis and results. Section 5 deals with comparative analysis and Section 6 presents the final conclusion.

STATE-OF-ART

Clustering Algorithms

Clustering is the process that identifies the natural groupings of data from a large data set. Clusters are formed in such a way that objects in the same cluster are similar and objects in different clusters are distinct. Group of similar objects is known as cluster. Similarity can be measure in various terms such as distance, membership value etc. depending on the various clustering techniques. The most widely used clustering algorithms are K-means [1] and Fuzzy C-means [2].

K-means partitioned data set into 'k' different subsets or clusters. It operates on actual data. It is centroid based clustering algorithm. It is usually similar to the Gaussian Mixture Method as both are iterative approaches and uses cluster centers to the model the data. Gaussian mixture models trained with EM (expectation-maximization) algorithm but K-means minimizes the sum of distances from each data vector to its cluster centers. In this algorithm, cluster centers are calculated as mean of all the data points assigned to the different clusters but Gaussian methods uses multivariate Gaussian distributions. Each data vector belongs to the cluster with the nearest cluster center but Gaussian mixture method uses probabilistic assignment to the clusters. It finds the cluster of the data points on the basis of minimum distance between center and data point. Initially, it consider random selection of data point as cluster center or mean of the data points or median of data points etc. and after that in each iteration, it computes cluster center and then distance between cluster center and data point to find their cluster. When the cluster center becomes constant then it stops iterating. Finally, it provides more compact and well separated clusters as a result because it minimizes the sum of distances from each data point to its cluster center, over all clusters.

Fuzzy c-means is one of the popular data clustering technique. In this technique, each data point belongs to a cluster to some degree that is specified by a membership value. It classifies data set on the basis of membership degree. It calculates membership degree using distance between center of the cluster and data point. This technique was originally introduced by Jim Bezdek in 1981 [2] as an improvement on earlier clustering methods.

There are various other clustering techniques such as genetic algorithm [21], SLINK (hierarchical clustering) [22], global k-means [23] etc.

Various Validity Indexes

In the literature, various validity indexes have been proposed to determine the optimal number of clusters.

This work proposed a validity index based on standard deviation and data dimensions to find optimal number of clusters. This section focuses on some of the most common existing validity indexes. They are summarized in the Table 1.

Table 1: Summary of Some of the Existing Validity Indexes

Validity Index	Formula	Best at
The Partition Coefficient [12]	$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2$	Maximum value
The Partition Entropy [13][14]	$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\log_a u_{ij}$	Minimum Value
Modified Partition Coefficient [15]	$V_{MPC} = 1-\frac{c}{c-1}\left(1-\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\right)$	Maximum value
Krzanowski - Lai Index [16]	$V_{KL(c)} = \left\|\frac{D(c)}{D(c+1)}\right\|$ where, $W(c) = \frac{1}{N}\sum_{i=1}^{N}x_i - C_i^2$ and $D(c) = (c-1)^{\frac{2}{p}}W(c-1) - (c)^{\frac{2}{p}}W(c)$	Maximum value

Validity Index	Formula	Best at
Silhouette (Sil) [17]	$V_{SC} = \frac{1}{N}\sum_{i=1}^{N} S(i)$ where, $S(i) = \frac{q(i) - p(i)}{\max\{p(i), q(i)\}}$	Maximum value
Calinski – Harabasz Index [18]	$V_{CH} = \frac{\left(\frac{SS(BG)}{c-1}\right)}{\left(\frac{SS(WG)}{N-c}\right)} = \frac{\left(\bar{d}^2 + \frac{N-c}{c-1} A_c\right)}{\left(\bar{d}^2 - A_c\right)}$	Maximum value
Davies-Bouldin Index [19]	$V_{DB} = \frac{1}{c}\sum_{i=1}^{c} D_i$ where, $D_i = \max_{i \neq j} p_{ij}$ and $p_{ij} = \frac{\sigma(C_i) - \sigma(C_j)}{C_i - C_j^2}$	Minimum value
Non Fuzzy Index [20]	$V_{NFI} = \frac{c(\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}) - N}{N(c-1)}$	Maximum value
Weighted inter to intra cluster ratio [24]	$V_{Wint} = \left(1 - \frac{2c}{N}\right).WR(c)$ $WR(c) = 1 - \frac{\sum_{i=1}^{c} \frac{N_i}{N - N_i} \sum_{j\{1,\ldots i-1,i+1,\ldots c\}} N.\,Inter(C_i, C_j)}{\sum_{i=1}^{c} Intra(C_i)}$	Maximum value
Dunn's Validity Index [25] [26]	$V_{DI(c)} = \min_{i \epsilon c}\left\{\min_{j \epsilon c, i \neq j}\left\{\frac{\min_{x \epsilon c_i, y \epsilon c_j} d(x,y)}{\max_{k \epsilon c}\{\max_{x,y \epsilon c} d(x,y)\}}\right\}\right\}$	Maximum value

Proposed Methodology

K-means Algorithm

K-means clustering algorithm [1] [36] is a very common approach for pattern recognition. It is a clustering technique that partition n data vectors into k clusters. K-means is an iterative refinement technique that minimizes the sum of distances from each data vector to its cluster centers. It is basically focus on the distance between cluster center and data vectors. The center for each cluster is the point to which the sum of distances from all data vectors in that cluster is minimized. Initially, cluster centers are choose as a random data vectors from the data set.

K-means can be formulated as

Minimize

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k} x_i - v_j^2 \qquad (1)$$

where, $X = \{x_1, x_2, x_3, \ldots x_n\}$ is a data set of real vectors, n is the number of data points, $V = \{v_1, v_2, \ldots, v_n\}$ is cluster centers and k is the number of clusters.

Distance and center can be calculated as follows:

$$D_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{k} x_i - v_j^2 \quad \text{and} \quad v_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{u_{ij} * x_j}{u_{ij}} \qquad (2)$$

$$v_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{u_{ij} * x_j}{u_{ij}} \qquad (3)$$

where, m is the number of features (or data vector dimension) and u is the partition matrix. Assignment can be done on the basis of minimum distance from the cluster center.

Step 1: Initialize 'k' cluster centers

(Choose random data points from data set)

Step 2: Calculate distance between data vectors and cluster centers using Eq. (2).

Step 3: Assign data points to the closest cluster center.

Step 4: Update cluster center based on the assignment using Eq. (3).

Step 5: Repeat step 2 to 4 until the cluster centers become constant.

Pseudo - Algorithm Description

The pseudo algorithm is divided into two phases. The first phase identifies number of clusters and the second improves classification accuracy.

Phase I: In this phase, first apply k-means on the normalize data and calculate proposed validity index using Eq. (10). This process is repeated until the validity index value is less than or equal to zero otherwise return previous number of clusters which is the optimal number of clusters.

Phase II: Fine clustering is done in this phase.

Chun Su [28] proposed SBKM algorithm in 2001. SBKM is two phase algorithm: first phase is coarse tuning (executing K-means) and fine tuning which calculate point symmetry distance to improve clustering result. It calculates cluster centroid nearest it in the symmetric sense. But it fails in the case of overlapped and more complex data.

Clustering accuracy is difficult in case of overlapped clusters. In most of the clustering algorithm, data points are assigned to the closest cluster center like k-means, FCM etc.. But sometimes, some data points have very minor difference between two nearest clusters. Assign data points to the closest cluster center but actually data point belongs to the second nearest cluster. Because of the minor difference data point is assigned to the wrong cluster. This is a common problem in many clustering algorithms especially in case of multi-dimensional, overlapped or complex datasets. If this problem can be rectified to some extent then clustering result must be improved automatically. This work solves this problem in phase II: fine clustering.

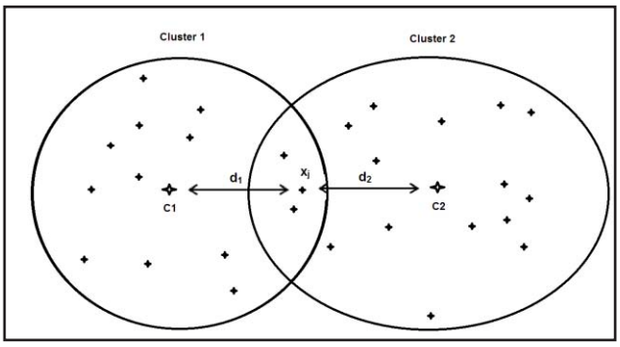


Fig. 1: An Example of Two Overlapped Clusters

Consider Fig. 1 as an example, where two overlapped clusters are shown with cluster centers C1 and C2. Take a data point x_j which comes under overlapped area between two clusters. Its distance from both the cluster centers C1 and C2 are d_1 and d_2 respectively. x_j is situated almost middle of the C1 and C2 but $d_1 < d_2$. So, x_j is assigned to the cluster 1 but actually it belongs to the cluster 2. In this situation, it is difficult to increase cluster accuracy. In this phase, this situation is trying to be solved at some extent. By updating two nearest cluster centers of a data point whose is the minimum in the set of all data points. First, calculate the difference between the two nearest cluster centers of the data points and then find the minimum difference m_d. In other words, find the data point which has minimum difference between two nearest cluster centers. m_d is defined as:

$$m_d = \text{minimum}\left\{\begin{matrix} |d_1 - d_2| \\ \textit{for all the data points} \end{matrix}\right\} \quad (4)$$

where, d1 is the distance between data points and center of the cluster 1 and d2 is the distance between data points and center of the cluster 2. Update these two nearest cluster centers by adding (to the second nearest) or subtracting (to the nearest) θ. θ is a small value because when update the cluster centers, they must belong to that cluster only. It can be calculated as

$$\theta = \sqrt{\frac{m_d}{2}} \quad (5)$$

It may be user-defined small value. After updating nearest cluster centers, calculate Euclidean distance using Eq. (2) and update partition matrix. Re-calculate cluster centers using Eq. (3). This process continues till cluster centers become constant. The classification accuracy was increased by repeating this process δ times to attain more accurate clustering. δ is user defined value. In this work, Assume δ = 10. Choose best out result of the ten iterations for the experiment. The different values of δ can be taken for different datasets to attain better accuracy results.

The steps of the pseudo algorithm are given below:

Step 1: Initialize c = 2

Step 2: Normalize Data Set

$$X_{ij} = X_{ij} - X_{min} / X_{max} - X_{min} \quad (6)$$

Step 3: Identify Number of Clusters

(Phase I)

Step 3.1: Loop

Step 3.2: Run K-means for 'c' clusters

Step 3.3: Calculate proposed validity index, V (u, λ)

Step 3.4: If V(u,) ≤ 0 then c = c + 1else terminate the loop.

The optimal number of clusters is (c – 1).

Step 4: Fine Clustering (Phase II)

Step 4.1: for I = 1 to δ

Step 4.2: Calculate the minimum difference (min_diff) from the distance between two nearest clusters of each data point.

Step 4.3: Update two nearest cluster centers using below given equation

$$v_{ij} = v_{ij} \pm \theta \tag{7}$$

Step 4.4: Calculate Euclidean distance using Eq. (2) from the new centers & update partition matrix.

Step 4.5: Calculate cluster centers using Eq. (3)

Step 4.6: If cluster center becomes constant then break for loop otherwise continue loop

Step 5: Display clustering result

Variance and Standard Deviation[39]

Variance is the measure of the variations between the values in a data set. It is denoted by σ^2. Standard deviation shows how much variation/dispersion exists from the mean value.

It is denoted by σ. The standard deviation is calculated as

$$\sigma = \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right\}^{\frac{1}{2}} \tag{8}$$

where, *n* is the number of elements in the sample, and *y* is the data vector.

A small standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

This concludes that when standard deviation of a cluster decreases more compact the cluster distribution.

Membership Value

The membership value between data points and cluster centers is calculated as

$$M_f = 1/\sum (D_{ij}/D_{ik})^2 \tag{9}$$

where, D is the matrix that stores distance between datasets and centers. i ϵ 1:M, j,k ϵ 1:c. c is the number of clusters and M is the number of data points. Details are available in Ref. [27].

Proposed Validity Index

Validity of the clusters can be measured in various ways. There are several validity indexes introduced in the literature. Some of them based on membership values, distance between data and cluster center such as PC, MPC, PE etc. Some of them based on variation measured in the clusters such as variance criteria given by Willian A. Porter, Davies-Bouldin Index, Liang-Zhang-Zhao's index (S) etc.

First consider V_{PC} [12] to explain proposed criteria

$$V_{PC} = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2 \tag{10}$$

Where u_{ij} is the membership value and V_{PC} is modified by subtracting mean membership value.

$$\frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij} - \overline{u_{ij}})^2 = variance = \sigma^2(u) \tag{11}$$

It is variance of membership values. Variance is used to check compactness of the cluster.

Consider V_{MPC}[15]

$$V_{MPC} = 1 - \frac{c}{c-1} \left(1 - \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2 \right) \tag{12}$$

It is based on PC. Data dimension and number of clusters used along with sum of variance of membership values in most of the validity indexes.

The proposed validity index is based on PC and MPC. It is defined as

$$V(u, \lambda) = 1 - \left(\frac{1}{\lambda} \right)^2 * \left(\sigma^2(u) \right) \tag{13}$$

$V(u, \lambda)$ is calculated for each cluster.

$\sigma^2(u)$ is the variance of membership value using Eq. (3) and λ is defined as

If M > N

$$\lambda = \frac{Nc^2}{M} \tag{14}$$

Otherwise,

$$\lambda = \frac{2Mc^2}{N} \tag{15}$$

Where, c is the assumed number of clusters, M is the number of data points and N is the number of features.

More compact the cluster lesser the variance [6] and Standard deviation ratios [24].

Validity indexes gives optimal number of clusters depends upon the minimum and maximum values of the index. This work proposed that V(u, λ) value is optimal when it is less than or equal to zero because minimum value of variance gives more compact clusters.

$$1 - \left(\frac{1}{\lambda} \right)^2 * \left(\sigma^2(u) \right) \leq 0 \tag{16}$$

Multiply equation by λ^2, we get

$$\lambda^2 - \left(\sigma^2(u)\right) \leq 0 \quad (17)$$

$$\left(\sigma^2(u)\right) \geq \lambda^2 \quad (18)$$

Taking square root on both sides, we get

$$\sigma(u) \geq \lambda \quad (19)$$

Now, it is a simplified form of V(u, λ). It can be used as validation criteria in the proposed algorithm to reduce calculations.

When the standard deviation exceeds λ then it means cluster number is higher than optimal cluster. So, (*c*–1) is the correct number of clusters for the data vector.

Initialize Cluster Centers

There are several cluster initialization methods available in the literature [27] [37] [38]. *Pena* [38] compares four different initialization methods for K-means that are RANDOM, Forgy approach, Macqueen approach and Kaufman approach. *Malinen* [37] explains seven different initialization methods for k-means. The most common and widely used initialization method of K-means is random selection method because of its simplicity. The initial clusters are selected randomly among the dataset. In this method, data vectors are initialized randomly to these cluster locations. Random selection may cause some undesirable situation such as sometimes the lower cardinality clusters can more easily become empty. The initialization methods help us to reduce multiple iterations of the algorithm and convergence speed [37] [38]. While using random selection method, multiple execution of the algorithm may be required to arrive at a meaningful conclusion.

Thus, this work uses a different initial cluster assignment method to show the fix experiment result, increases speed and undesirable conditions such as empty cluster.

The assignment can be done in the following manner.

Choose k data points from the data set as initial cluster centers. These data points were selected as arithmetic progression series which starts with 1st point and followed by the difference (M/k). Where, M is the number of data points and k is the number of assumed clusters.

1^{st}, $(1+(M/k))^{th}$, $(1+2(M/k))^{th}$, ……. $(1+(k-1)(M/k))^{th}$

Above series is considered as positions of the data points which are used as initial cluster centers.

For example,

If there are 100 data points wants to be divided into three clusters. So, M=100 and k=3.

1^{st}, $(1+(100/3)) = (1+33)=34^{th}$ and

$(1+(3-1)(100/3)) = (1+2*33) = (1+66)=67^{th}$

data points are considered as initial cluster centers.

Various other cluster initialization methods like random selection can be used to initialize cluster centers for the proposed pseudo-algorithm.

Lambda Selection

In the proposed work, lambda is defined as $\frac{Nc^2}{M} if\ M > N\ otherwise \frac{2Mc^2}{N}$ where N is the number of features, M is the number of data points and c is the assumed number of clusters. This work tested different combinations of data dimension and assumed number of clusters for lambda. The comparison of result is shown in Table 2.

Table 2: Number of Clusters Identified by the Proposed Pseudo-algorithm using Different Lambda

Dataset	Actual No. of Clusters	$if\ M > N\lambda = \frac{Nc^2}{M} else \frac{2Mc^2}{N}$	$if\ M > N\lambda = \frac{c^2(N+c-1)}{M}$	$else \frac{c^2(M+c-1)}{N}$	$if\ M > N\lambda = \frac{c^3}{M} else \frac{c^3}{N}$	$if\ M > N\lambda = \frac{\lambda = 1}{Nc^{c-1}} else \frac{Mc^{c-1}}{(M+N)}$	$\lambda = \frac{c^c}{(M+N)}$
Iris	3	3	3	3	3	3	3
Thyroid	3	3	2*	3	3	3	3

WDBC	2	2	2	4*	2	3*	4*
D2C3	3	3	2*	3	3	3	3
D3C4	4	4	3*	4	3*	3*	4
D2C5	5	5	4*	4*	3*	3*	4*
D2C6	6	6	4*	4*	3*	3*	4*

*wrong identification

Experimental Analysis and Results

In this section, the performance of the both the phases of the pseudo-algorithmwas evaluated and Phase I compare with various validity indexes and Phase II compares with K-Means and SBKM by applying it to both real and synthetic datasets.

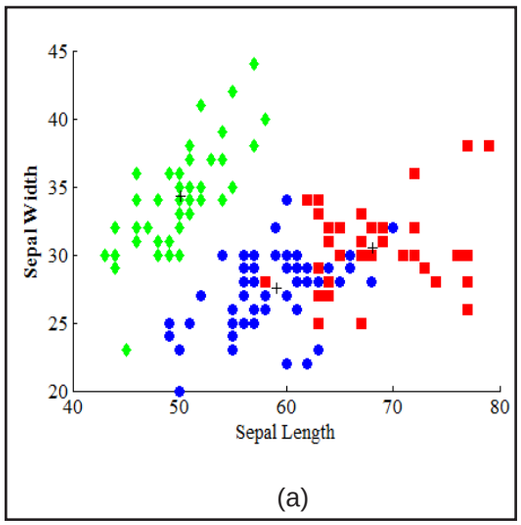
Clustering Result of Real Datasets

Data 1: IRIS data set is the most popular dataset. The dataset contains 150 data points with four features. These four features are sepal length, sepal width, petal length and petal width. This database contains three classes of data points named Virginica, Setosa and Veriscolor. Each class contains 50 data points. Class II and Class III are overlapped classes.

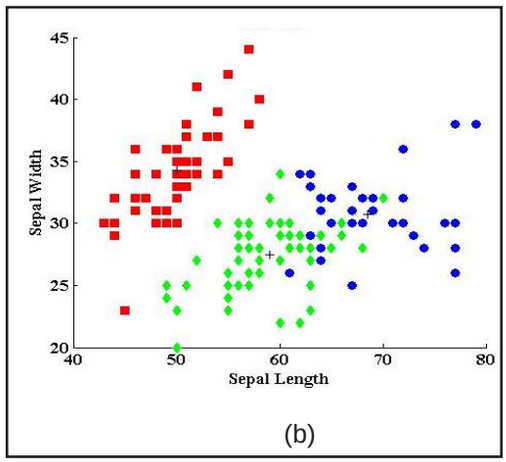
The proposed work gives best result with 91.33% accuracy. Both K-Means and FCM attain 88.67% accuracy, FRBC gives 88.67% accuracy and SBKM classify dataset into two clusters only. Both k-means and FCM also gives good results. Classification results are shown in Fig. 2 and Classification result compared with FRBC and FCM is shown in Table 3.

Table 3: Comparison of Clustering Result of Proposed Work with FRBC and FCM for Iris Dataset

Actual Class Label	Proposed work			FRBC [9]			FCM [9]		
	Predicted Class Label			Predicted Class Label			Predicted Class Label		
	1	2	3	1	2	3	1	2	3
1	50			50			50		
2		49	01		50			46	4
3		12	38		17	33		12	38



(a)



(b)

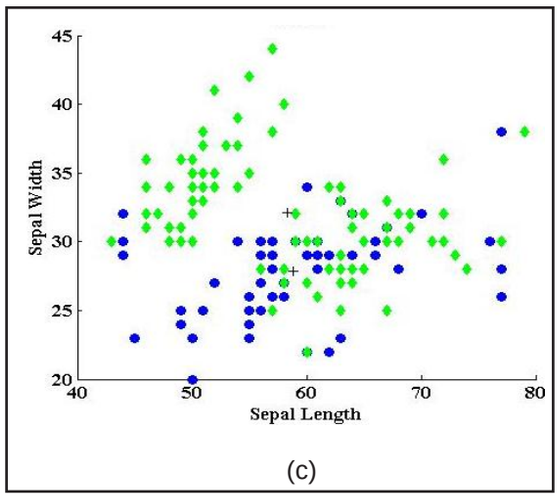


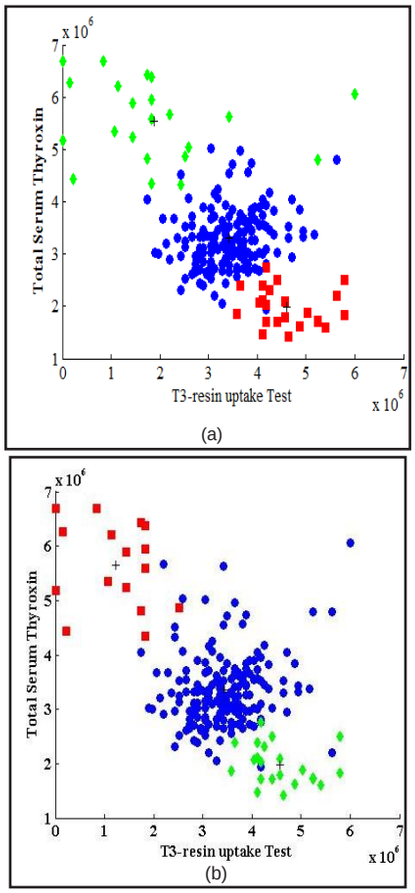
Fig. 2: Classification Result of IRIS Dataset using (a) Pseudo-algorithm (b) K-Means (c) SBKM

Data 2: Thyroid database contains 215 data points. Each data point has five features. The results of the tests are considered as classes. There are three classes that are normal, hyper and hypo. Normal class has 150 instances, hyper has 35 and 30 instances are of hypo class.

Proposed work gives best result with 90.7% accuracy. K-Means gives 88.84% accuracy and SBKM, FCM and FRBC classify dataset into two clusters only. Classification results are shown in Fig. 3 and Classification result compared with FRBC and FCM is shown in Table 4.

Table 4: Comparison of Clustering Result of Proposed Work with FRBC and FCM for Thyroid Dataset

Actual Class Label	Proposed work			FRBC [9]			FCM [9]		
	Predicted Class Label			Predicted Class Label			Predicted Class Label		
	1	2	3	1	2	3	1	2	3
1	150			150			143	7	
2	13	22		12	23			35	
3	07		23	30			30		



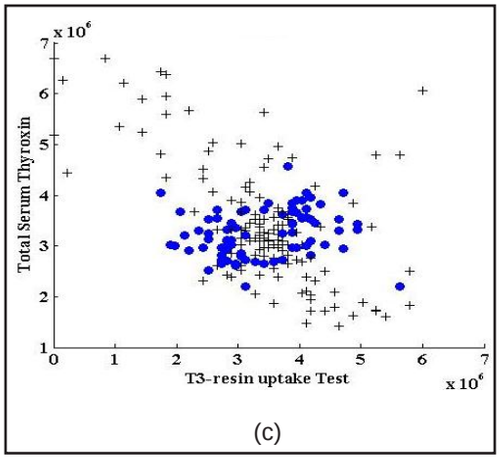


Fig. 3: Classification Result of Thyroid Dataset using (a) Pseudo-algorithm (b) K-Means (c) SBKM

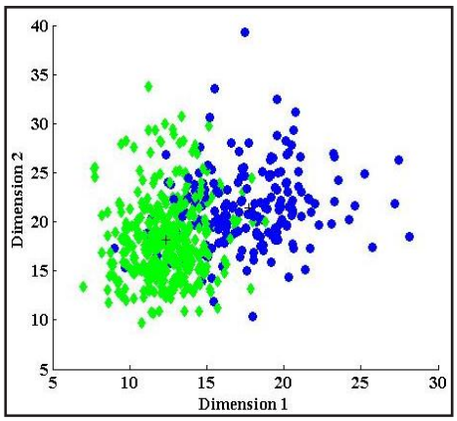
Data 3: Wisconsin Diagonistic Breast Cancer (WDBC) dataset contains 569 data points with 30 features. This dataset contains data of two different classes. 212 data points are from one class and 357 data points from the another class.

The proposed work gives best result with 93.32% accuracy. K-Means attain 92.79% accuracy, FCM gives 92.79% accuracy, FRBC attains 91.21% accuracy and SBKM classify dataset with 52.90% accuracy. FCM,

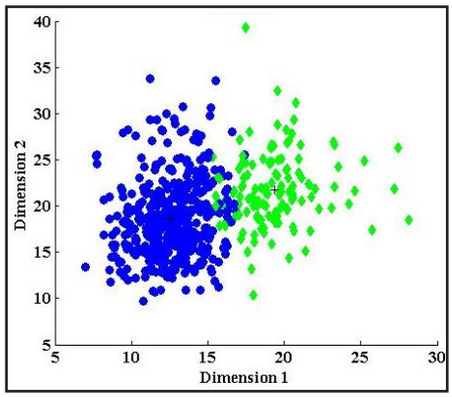
FRBC and proposed work gives more than 90% accuracy. Classification results are shown in Fig. 4 and Classification result compared with FRBC and FCM is shown in Table 5.

Table 5: Comparison of Clustering Result of Proposed Work with FRBC and FCM for WDBC Dataset

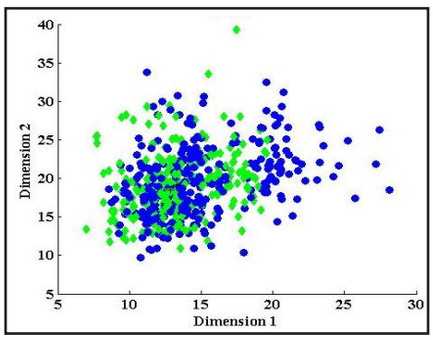
Actual Class Label	Proposed work		FRBC [9]		FCM [9]	
	Predicted Class Label		Predicted Class Label		Predicted Class Label	
	1	2	1	2	1	2
1	180	32	45	167	28	184
2	06	351	352	05	344	13



(a)



(b)



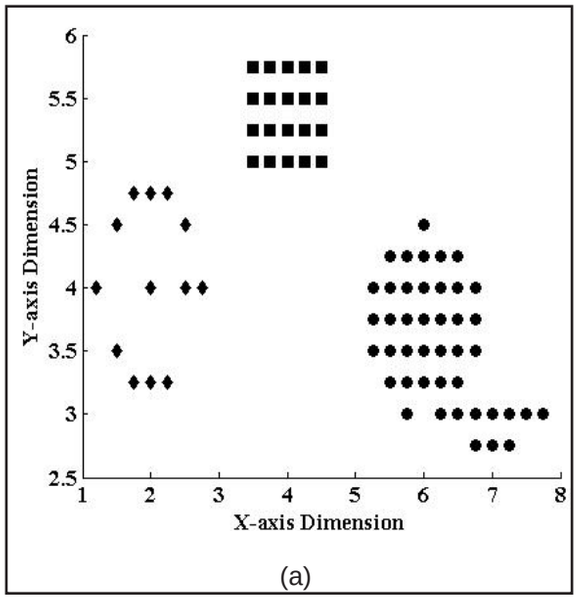
(c)

Fig. 4: Classification Result of WDBC Dataset using (a) Pseudo-algorithm (b) K-Means (c) SBKM

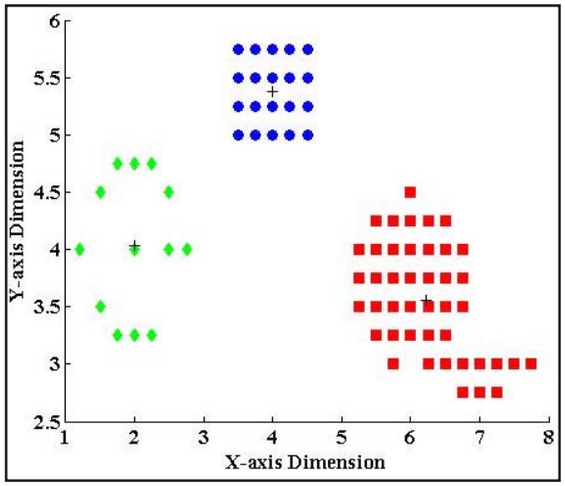
Clasification Result of Synthetic Datasets

Data 4: D2C3 is the artificial dataset. It contains 76 datapoints with two dimensions. This data set contains three different shapes that are outlined small letter 'e', filled 'Q' and filled rectangle. Each shape represents a cluster. So, the data set is divided into three separated clusters.

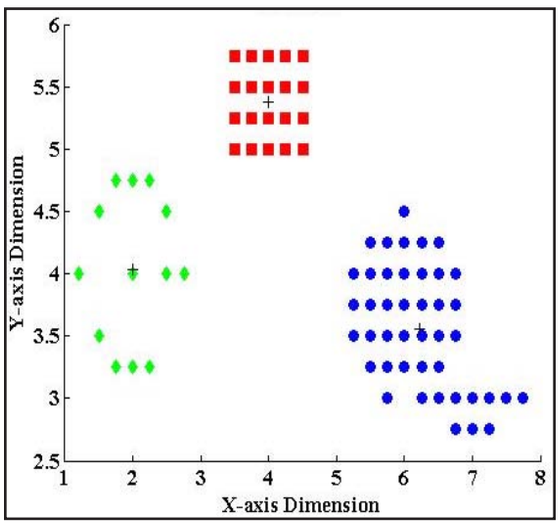
Both K-Means and proposed work classify dataset with 100% accuracy shown in Fig. 5 and confusion matrix in Table 6. SBKM attains 84.21% accuracy.

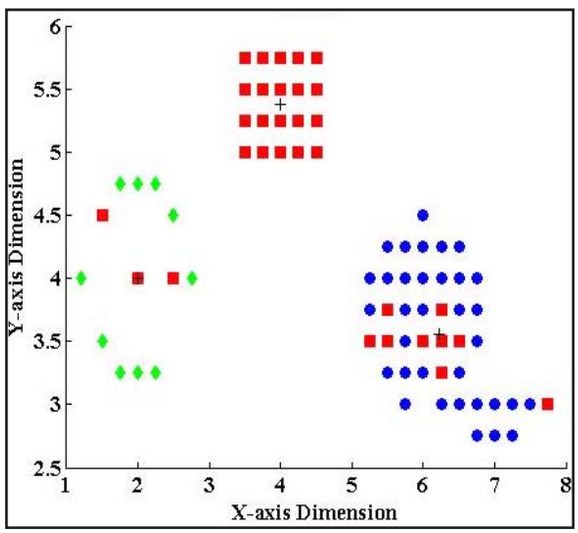


(a)



(b)





(d)

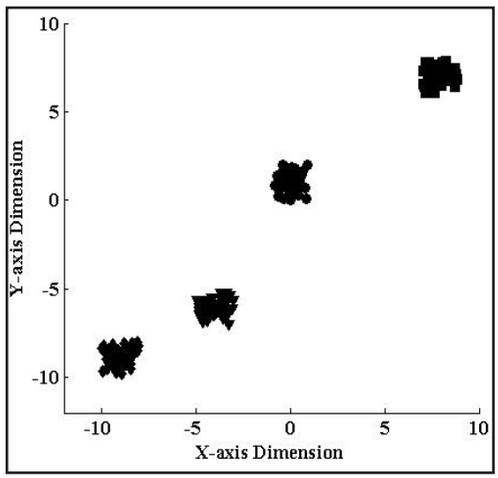
Fig. 5: (a) D2C3 Dataset; Classification Result of D2C3 Dataset using (b) Pseudo-algorithm (c) K-Means (d) SBKM

Table 6: Confusion matrix of D2C3 Dataset

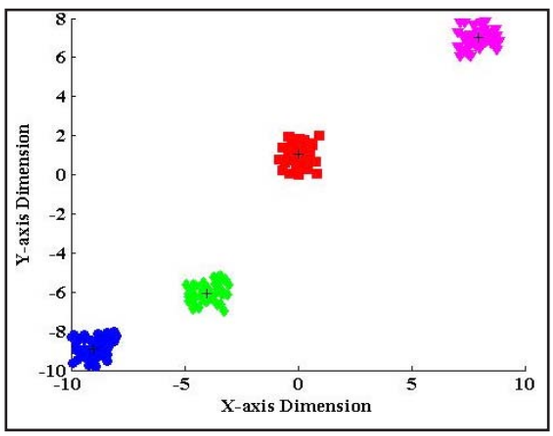
Actual Class → Predicted Class ↓	Class I	Class II	Class III
Class I	13	0	0
Class II	0	43	0
Class III	0	0	20

Data 5: D3C4 is the artificial dataset which has three dimensions and four clusters. It contains 200 datapoints. All four clusters are compact & well-seperated.

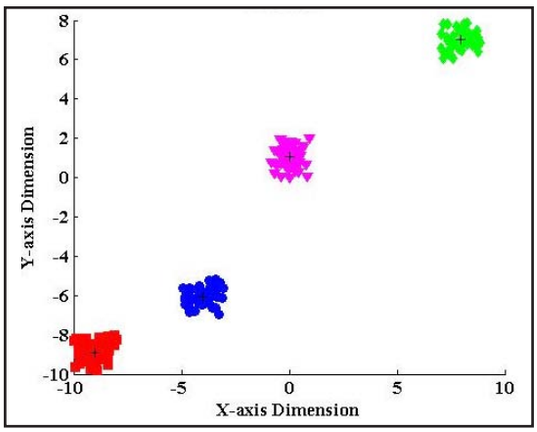
The proposed work and K-Means classify dataset with 100% accuracy but SBKM classifies into only two clusters shown in Fig. 6 and confusion matrix in Table 7.



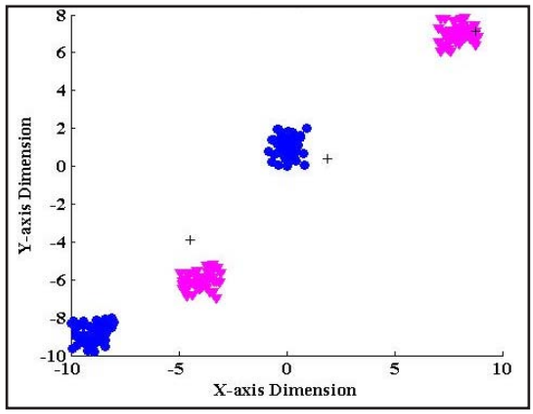
(a)



(b)



(c)



(d)

Fig. 6: (a) D3C4 Dataset; Classification Result of D3C4 Dataset using (b) Pseudo-algorithm (c) K-means (d) SBKM

Table 7: Confusion Matrix of D3C4 Dataset

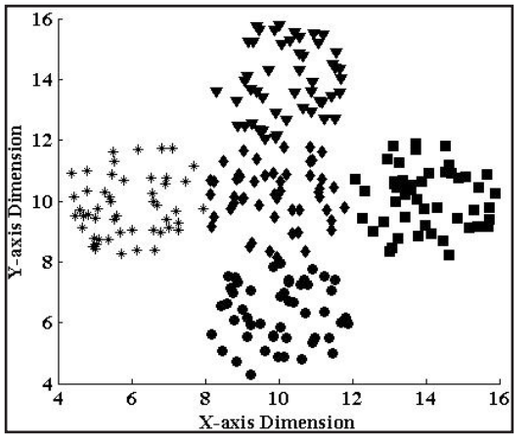
Actual Class → Predicted Class ↓	Class I	Class II	Class III	Class IV
Class I	50	0	0	0
Class II	0	50	0	0
Class III	0	0	50	0
Class IV	0	0	0	50

Data 6: D2C5 is the two-dimensional artificial dataset. This datset is divided into five spherical clusters. Each cluster contains 50 data points. There are 250 data points in this dataset.

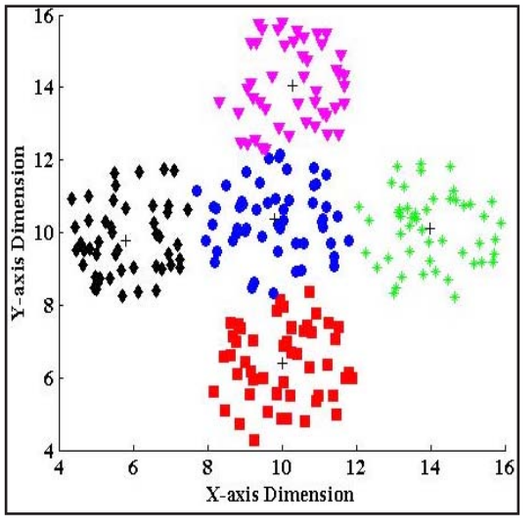
Both proposed work and K-Means attains 100% classification accuracy but SBKM classifies into four clusters shown in Fig. 7 and confusion matrix in Table 8.

Table 8: Confusion Matrix of D2C5 Dataset

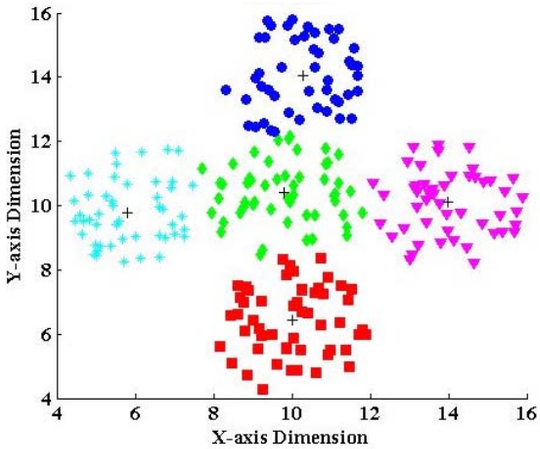
Actual Class → Predicted Class ↓	Class I	Class II	Class III	Class IV	Class V
Class I	48	2	0	0	0
Class II	0	48	2	0	0
Class III	0	0	50	0	0
Class IV	0	3	0	47	0
Class V	0	0	0	0	50



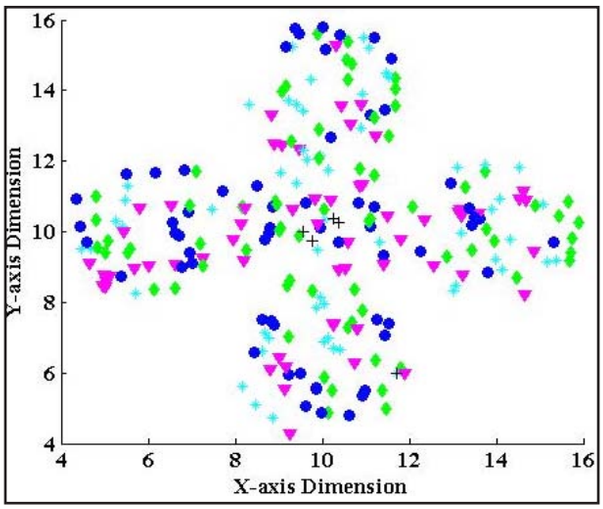
(a)



(b)



(c)



(d)

Fig. 7: (a) D2C5 Dataset; Classification Result of D2C5 Dataset using (b) Pseudo-algorithm

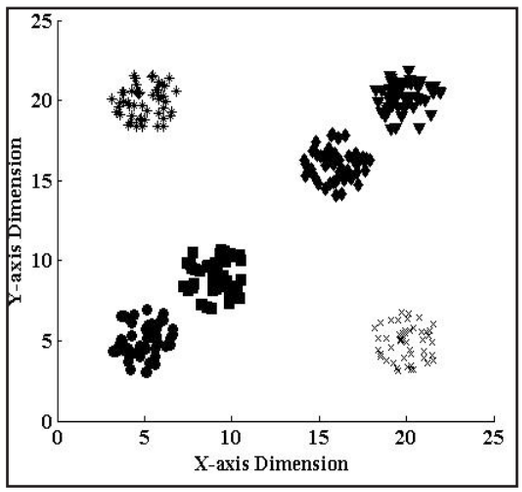
(c) K-Means (d) SBKM

Data 7: D2C6 is also two-dimensional artificial dataset. There are 300 data points in this dataset. This datset is divided into six spherical clusters.

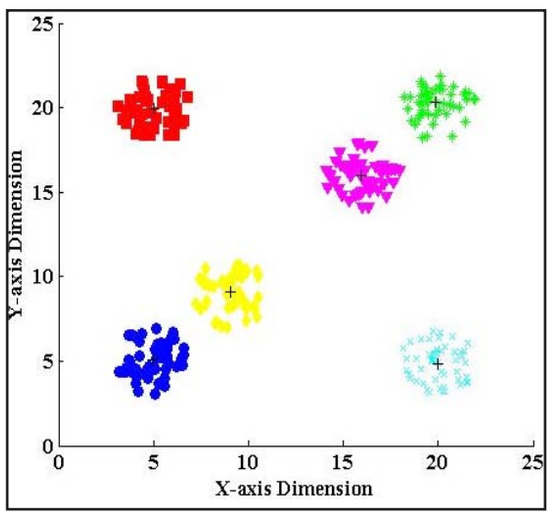
K-Means and Proposed work classify dataset with 100% accuracy except SBKM shown in Fig. 8 and confusion matrix in Table 9. SBKM classifies into four clusters.

Table 9: Confusion matrix of D2D6 Dataset

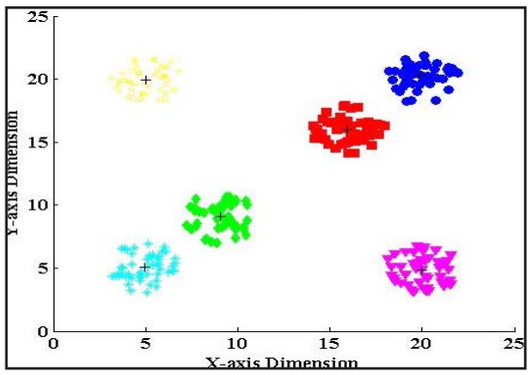
Actual Class → Predicted Class ↓	Class I	Class II	Class III	Class IV	Class V	Class VI
Class I	50	0	0	0	0	0
Class II	0	50	0	0	0	0
Class III	0	0	50	0	0	0
Class IV	0	0	0	50	0	0
Class V	0	0	0	0	50	0
Class VI	0	0	0	0	0	50



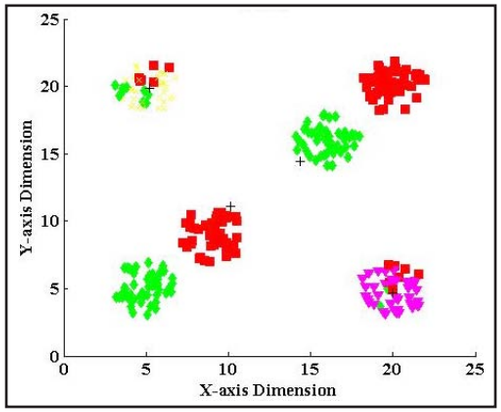
(a)



(b)



(c)



(d)

Fig. 8: (a) D2C6 Dataset; Classification Result of D2C6 Dataset using (b) Pseudo-algorithm (c) K-Means (d) SBKM

5. COMPARATIVE ANALYSIS

Comparing Proposed Validity Index with other Validity Indexes to Find Optimal Number of Clusters

The previous proposed validity indexes explained in section 2 gives result as minimum or maximum value after calculating index value for particular range of the given number of clusters. But, the proposed validity index does not calculate its result on the basis of either minimum or maximum value of the index. Thus, it can't be calculated for a particular range of assumed number of clusters. It is calculated till its value is positive. While its value is either zero or negative, it means the previous assumed number of clusters is the optimal number of clusters. This work concluded that it is faster than other proposed validity indexes especially in those cases where number of clusters is not large. The various validity indexes used to compare with the proposed validity index are listed in Table 10.

In IRIS dataset, there are three clusters as discussed in section 4.1. Two of them are overlapped. Some of the literature papers [34] [35] consider both two clusters and three clusters are optimal. This work clearly identifies three clusters.

2 to N-1 is the best range to identify correct result, where n is the number of data points. This range may be large in some cases. It is difficult to represent all values here in those cases. Thus, 2 to 10 range was used to calculate values of other validity indexes. The number of clusters identified by the proposed validity index compared with other validity indexes is shown in Table 10.

Comparative Results of Classification

Accuracy rate of pseudo-algorithm is compared with K-Means, SBKM [28], FCM and FRBC [9]algorithms are shown in Table 11.

Table 10: Comparisons of Proposed Validity Index with Other Validity Indexes

Dataset	Actual No. of Clusters	Proposed Validity Index	PC	MPC	PE	NFI	Sil	DB	CH	KL	Wint	Dunn	OS
Iris	2/3	3	2	2	2	2	2	2	3	3	3	2	2[34]
Thyroid	3	3	2*	2*	2*	2*	2*	2*	3	3	2*	2*	-
WDBC	2	2	2	2	2	2	2	2	2	9*	9*	2	2[35]

D2C3	3	3	3	3	3	3	3	3	6*	9*	3	3	-
D3C4	4	4	4	4	4	4	4	4	4	4	3*	4	-
D2C5	5	5	5	5	5	5	5	5	5	2*	3*	5	-
D2C6	6	6	6	6	6	6	4*	6	6	8*	4*	6	-

*Wrong Identification

Table 11: Comparisons of Classification Accuracy Rate of Proposed Algorithm with K-means and SBKM

Dataset	Proposed Work	K-Means	SBKM	FRBC	FCM
Iris	91.33%	88.67%	59.33%*	88.67%	89.33%
Thyroid	90.70%	88.84%	50.23%*	80.47%*	82.79%*
WDBC	93.32%	92.79%	52.90%	91.21%	92.79%
D2C3	100.00%	100.00%	84.21%	-	100%
D3C4	100.00%	100.00%	25.00%*	-	100%
D2C5	97.20%	96.80%	19.60%#	-	94.80%
D2C6	100.00%	100.00%	16.67%#	-	100%

Classify into* two clusters #four clusters only.

Conclusions

The k-means is a most widely used clustering algorithm. K-means is an unsupervised clustering algorithm. It requires number of clusters in advance. The proposed work overcome this problem by applying a new validity index based on standard deviation and data dimension to determine the optimal number of clusters and after that apply modified k-means clustering to get better clustering results. This approach gives us a way to classify datasets accurately without pre-specifying number of clusters. Pseudo-algorithm has two phases: one identifies correct number of clusters and second increases classification accuracy.

There are lots of validity indexes available in the literature to identify number of clusters. Most of the available validity indexes such as Davies-Bouldin[19], Calinski-Harabasz[18], Partition Coefficient [12], Partition Entropy [13][14], Silhouette [17], Modified Partition Coefficient [15], Krzanowski& Lai [16], Non Fuzzy Index [20] etc. calculate its value for the given range of number of clusters and finally, identify the actual number of clusters depending on the minimum or maximum value of the validity index. They fail if actual number of clusters is higher than the specified range of clusters. The proposed validity index does not require range of number of clusters. It starts calculated for two clusters and increased by one every time up to the actual number of clusters. In this way, it reduces the unnecessary calculation and execution time.

The pseudo-algorithm requires only dataset as input and returns clustering result (class labels). In this way, it automatically identifies number of clusters. The proposed work first identifies number of clusters using proposed validity index and then continues classify the dataset into the actual number of clusters by applying proposed modified k-means. That is why; it is a two phase algorithm.

Confusion matrix is used to identify the classification accuracy.

Experiments performed on three real datasets and four artificial datasets proved that pseudo algorithm not only reduces classification errors but also identifies number of clusters automatically.

In future, both the phases can be used separately and will be extended (if required).

References

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1, 281-296.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). *FCM: The Fuzzy C-means Clustering Algorithm, Computers and Geosciences*, 10(23), 191-203, Pergamon Press Ltd., USA.

Sun, H., Wang, S., & Jiang, Q. (2004). FCM-based Model Selection Algorithms for Determining the Number

of Clusters, Elsevier, Pattern Recognition, (pp. 2027-2037).

Liang, Z., Zhang, P., & Zhao, J. (2010). Optimization of the Number of Clusters in Fuzzy Clustering. *IEEE International Conference on Computer, Design and Applications*.

Yan, Z., & Pi, D. (2009). A fuzzy clustering algorithm based on K-means. *International Conference on Electronics and Business Intelligence*.

Li, M. J., Ng, M. K., Cheung, Y. M., & Huang, J. Z. (2008). Agglomerative Fuzzy K-means Clustering algorithm with selection of number of clusters. *IEEE Transactions Knowledge and Data Engineering,* 20(11).

Xie, J., & Jiang, S. (2010). A simple and fast algorithm for global K-means clustering. *IEEE 2nd International Workshop on Education Technology and Computer Science*.

Bagirov, A. M. (2008). Modified global k-means algorithm for minimum sum-of-squares clustering problems, *Elsevier, Pattern Recognition*.

Mansoori, E. G. (2011). FRBC: A fuzzy rule-based clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 19(5).

Ben, S., & Su, G. (2009). A new validity index for fuzzy clustering. *IEEE Chinese Conference on Pattern Recognition*.

Wu, K. L., Yang, Y. M. S., & Hsieh, J. N. (2009). Robust cluster validity indexes. *Elsevier Journal on Pattern Recognition*, November, 42(11), 2541-2550.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function*, Plenum Press, New York.

Bezdek, J. C. (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3), 58-73.

Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1(1), 57-71.

Dave, R. N. (1996). Validating fuzzy partition obtained through c-shells clustering. *Pattern Recognition Letters*, May, 17(6), 613-623.

Krzanowski, W. J., & Lai, Y. T. (1988). *A criterion for determining the number of groups in a data set using sum of squares clustering*. *Biometrics*, 44, 23-34.

Rousseau, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1), 53-65.

Calinski, T., & Harabasz, J. (1974). A dendrite method of cluster analysis. *Communications in Statistics- Theory & Methods*, 3(1)1, 1-27.

Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(4), 224-227.

Anuar, N., & Zakaria, Z. (2010). Cluster Validity Analysis for Electricity Load Profiling. *IEEE International Conference on Power and Energy (PECon2010)*, Kuala Lumpur, Malaysia.

Goldberg, D. (1989). Genetic Algorithms in Search, Optimization and Machine Learning, *Addison-Wesley Longman Publishing Co.*, Inc., New York, NY, USA.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society)*, 16(1), 30-34.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2), 451-461.

Strehl, A., & Ghosh, J. (2002). Relationship-based Clustering and Visualization for High-dimensional Data Mining. *INFORMS Journal on Computing*, 15(2), 208-230.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, 3(3), 32-57.

Dunn, J. C. (1973). A fuzzy relative of the Isodata process and its use in detecting compact well-separated clusters. *Journal on Cybernetics*, 3(3), 32-57.

Sharma, J., Panchariya, P. C., & Purohit, G. N. (2013). Clustering Algorithm based on K-means and Fuzzy Entropy for E-nose Applications. *IEEE Conference on Advanced Electronics Systems*.

Su, M. C., & Chou, C. H. (2001). A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence***,** 23(6), 674-680**.**

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over Likelihood Ratio Tests. *Systematic Biology*, 53(5), 793-808.

Corduneanu, A., & Bishop, C. M. (2001). *Variational Bayesian Model Selection for Mixture Distributions, Artificial Intelligence and Statistics*, T. Jaakkola and T. Richardson (Eds), pp. 27-34.

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473), 168-178.

Kearns, M., Masour, Y., Ng, A. Y., & Ron, D. (1997). An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27(1), 7-50.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486-494.

Kim, D. W., Lee, K. H., & Lee, D. (2004). On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37(10), 2009-2025.

Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158(19), 2095-2117.

Hartigan, J. A., & Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1), 100-108.

Malinen, M. I., Istodor, R. M., & Fränti, P. (2014). K-means: Clustering by gradual data transformation. *Pattern Recognition*, 47, 3376-3386.

Pena, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20(10), 1027-1040.

Pfeiffer, P. E. (1990). Variance and standard deviation, Chapter- Probability of Applications, Part of the series Springer Texts in Statistics, (pp. 355-370).