Int J CARS (2016) 11:2033–2047
DOI 10.1007/s11548-016-1437-9

ORIGINAL ARTICLE

Analysis of k-means clustering approach on the breast cancer Wisconsin dataset

Ashutosh Kumar Dubey[1] · **Umesh Gupta**[1] · **Sonal Jain**[1]

Received: 15 February 2016 / Accepted: 27 May 2016 / Published online: 16 June 2016

Abstract
Purpose Breast cancer is one of the most common cancers found worldwide and most frequently found in women. An early detection of breast cancer provides the possibility of its cure; therefore, a large number of studies are currently going on to identify methods that can detect breast cancer in its early stages. This study was aimed to find the effects of k-means clustering algorithm with different computation measures like centroid, distance, split method, epoch, attribute, and iteration and to carefully consider and identify the combination of measures that has potential of highly accurate clustering accuracy.
Methods K-means algorithm was used to evaluate the impact of clustering using centroid initialization, distance measures, and split methods. The experiments were performed using breast cancer Wisconsin (BCW) diagnostic dataset. Foggy and random centroids were used for the centroid initialization. In foggy centroid, based on random values, the first centroid was calculated. For random centroid, the initial centroid was considered as (0, 0).
Results The results were obtained by employing k-means algorithm and are discussed with different cases considering variable parameters. The calculations were based on the centroid (foggy/random), distance (Euclidean/Manhattan/Pearson), split (simple/variance), threshold (constant epoch/same centroid), attribute (2–9), and iteration (4–10). Approximately, 92 % average positive prediction accuracy was obtained with this approach. Better results were found for the same centroid and the highest variance. The results achieved using Euclidean and Manhattan were better than the Pearson correlation.
Conclusions The findings of this work provided extensive understanding of the computational parameters that can be used with k-means. The results indicated that k-means has a potential to classify BCW dataset.

Keywords Breast cancer · Breast cancer Wisconsin (BCW) diagnostic dataset · K-means · Foggy and random centroid

Introduction

Breast cancer, the second most common cancer across the world after lung cancer, is by far the most frequent cause of cancer death in women [1,2]. If it is diagnosed in early stages, the possibilities of survival are higher [3]. Since its symptoms vary from patient-to-patient, it is essential to characterize distinctive features of different patients and design a patient-specific treatment. The detection of the pattern of symptoms using data mining is a very important technique to correctly understand hidden patterns. The pertinent patterns extraction from the huge database is possible because of data mining techniques [4]. According to Jain et al. [5], data mining can be used for classification, estimation, prediction, association rules, clustering, and visualization activities. Of these activities, prediction, classification, and estimation come in supervised learning categories that prepare the model based on the available data representing one or more attributes. In these techniques, clustering is an important activity that enables grouping of data based on the nature or a symptom of the disease. So it can be applied in the primary stage for data pruning. K-means algorithm is one of the simple and important clustering algorithms. Classification

✉ Ashutosh Kumar Dubey
ashutoshdubey123@gmail.com

[1] JK Lakshmipat University, Near Mahindra SEZ, P.O. Mahapura Ajmer Road, Jaipur, Rajasthan 302 026, India

can be based on parametric, semi-parametric, or nonparametric approaches. The parametric approach uses sample from known distribution, nonparametric uses sample from an unknown distribution, while semi-parametric uses samples from both known and unknown distribution [6]. K-means clustering falls under semi-parametric approach, and it is an easier way of classifying dataset assuming k clusters. The main advantage of k-means is that it can have high computational speed for the large variable if the number of clusters is small. Bradley et al. [7] used k-means algorithm for the refinement of the initial points and achieved acceptable low run time. Similarly, Mary et al. [8] applied k-means algorithm to refine groups and applied ant colony optimization (ACO) for improving cluster quality.

In 2014, based on optimization process, molecular regularized consensus patient stratification (MRCPS) method of clustering was developed by Wang et al. [9] which has the capability to cluster both numerical and categorical data. Further, Rahideh et al. [10], based on k-means and fuzzy c-means, presented a classification approach, which is more accurate, sensitive, and specific than non-clustering method. K-means and fuzzy c-means methods have been used for clustering breast cancer data, and better memory, process time, and fitness points were achieved [11]. In addition, biased random-key genetic data clustering algorithm has been proposed by Festa et al. [12], which is relatively useful than other related methods. Chen et al. [13] proposed a hybrid intelligent model that is efficient in feature selection which has been used to analyze the clinical breast cancer data. Wei et al. [14] proposed a novel clustering algorithm for DNA sequence classification and their relationship using a novel clustering algorithm. Ahmad et al. [15] successfully designed a new k-means clustering algorithm that can use mixed numeric and categorical features. It was found to be effective in comparison with other clustering algorithms.

This study was undertaken to check the performance accuracy of k-means clustering algorithms on the breast cancer Wisconsin (BCW) diagnostic dataset. We studied following parameters:

- Accuracy of clustering in separating benign and malignant tumors.
- The effect of centroid, distance and splitting measures on k-means.
- The role of iterations and epoch in finding the centroid.
- The impact of the attributes of the BCW dataset on application of k-means clustering algorithm.

Materials and methods

BCW dataset from the University of Wisconsin Hospital was used to evaluate the impact of k-means clustering using centroid initialization, distance measures, threshold, attribute variations, and split methods. The objective of clustering this dataset was to achieve higher accuracy in cancer diagnosis. The principal aim of this algorithm is to expound k centers, one for each cluster. These focuses were set logically to avoid distinctive results from a diverse area selection. Along these steps, the better decision is to place them, however, much as could be expected far from one another. Every point was fitting in the given information set and relates to the closest center. The feature of this dataset is computed from fine needle aspirate (FNA) of a breast mass [16]. A total 699 instances, each having nine input attributes (2–10) and one target attribute (11) which is either benign or malignant, are presented in Table 1.

Table 1 Attribute information [16]

S. no.	Attribute	Domain
1	Sample code number (SCN)	Id number
2	Clump thickness (CT)	1 – 10
3	Uniformity of cell size (UCS)	1 – 10
4	Uniformity of cell shape (UCSh)	1 – 10
5	Marginal adhesion (MA)	1 – 10
6	Single epithelial cell size (SECS)	1 – 10
7	Bare nuclei (BN)	1 – 10
8	Bland chromatin (BC)	1 – 10
9	Normal nucleoli (NN)	1 – 10
10	Mitoses	1 – 10
11	Class	Benign (2), Malignant (4)

Notation used in Table 1

- Clump thickness (CT): It indicates grouping of cancer cells in multilayer.
- Uniformity of cell size (UCS): It indicates metastasis to lymph nodes.
- Uniformity of cell shapes (UCSh): It identifies cancerous cells, which are of varying size.
- Marginal adhesion (MA): It suggests loss of adhesion, i.e., a sign of malignancy but the cancerous cells lose this property. So this retention of adhesion is an indication of malignancy.
- Single epithelial cell size (SECS): If the SECS become larger, it may be a malignant cell.
- Bare nuclei (BN): Bare nuclei without cytoplasm coating which are found in benign tumors.
- Bland chromatin (BC): It usually found in benign cell.
- Normal nucleoli (NN): It is generally very small in benign cells.
- Mitoses: It is the process in cell division by which nucleus divides.

Notation used in algorithm

- $P1$: First random number.
- $P2$: Second random number.
- *NOR*: Number of rows.
- $K1$: First k-means.
- $K2$: Second k-means.
- *CS1*: Initial centroid 1 from the first epoch.
- *CS2*: Initial centroid 2 from the first epoch.
- *CL1*: Final centroid 1 from the last epoch.
- *CL2*: Final centroid 2 from the last epoch.
- d: Distance between two points.
- *Xi, Yi*: Coordinated of data points and cluster centers.
- n: Number of variables.
- N: Number of pairs.
- $\sum XY$: Sum of products of paired scores.
- $\sum X$: Sum of X scores.
- $\sum Y$: Sum of Y scores.
- $\sqrt{\sum X^2}$: Sum of squared X scores.
- $\sqrt{\sum Y^2}$: Sum of squared Y scores.
- PL: Partitions length.
- getRecords(): Function for fetching record.
- size(): It shows the number of array element.
- SL: Split records.

The basic idea behind the below algorithm is to efficiently use the k-means algorithm with different computation measures. For this BCW dataset has been selected with two cluster: one for malignant and another for benign. Initially, two centroid are needed as the number of cluster is 2. It is calculated based on foggy and random centroid according to the algorithm step 3. The distance between the cluster centers and the data point was measured using Euclidean, Manhattan, and Pearson coefficient according to the algorithm step 4. The data points with minimum distance from the cluster centers were assigned to the respective cluster. The splitting of data has been performed according to step 6. Mean and variance are calculated according to step 7 and step 8. Finally, new centroid is obtained based on the formula shown in step 9.

Algorithm: K-Means with the foggy and random centroid

Input: Input BCW dataset

Step 1: The number of clusters selected were k=2.

Step 2: The initial partition was positioned to classify the data into k clusters.

Step 3: K points were selected for starting assessments of the cluster centroids (These were the initial starting values). These two approaches were used for centroid initialization.

3.1 Foggy centroid

3.2 Two random numbers were found using the Java random function.

double P1= Math.random ();
double P2= Math.random ();
int Y1 = P1 * (NOR-1) +1;
int Y2 = P2 * (NOR-1) +1;
Y1 and Y2 rows were selected for the starting assessments of cluster centroids.

3.3 The centroids were chosen based on the selected attribute.

3.4 The centroid was then updated according to the cluster. The next attribute value was added to update the centroid.

3.5 The process 3.4 was continued until the epoch-1was reached or unless the same centroid was achieved.

3.6 Random centroid.

3.6.1 Initial centroid was considered as (0, 0).

3.6.2 A cluster was randomly assigned to each observation for computing of means.

3.6.3 The centroid was then updated based on the cluster. The next attribute value was added to make an updated centroid.

3.6.4 The process 3.6.3 was continued until the epoch-1 was reached or the same centroid was achieved.

Step 4: The distance between the cluster centers and the data point was measured using Euclidean, Manhattan, and Pearson coefficient.

4.1. Euclidean distance was computed by the following formula:

$d= \sqrt{\sum_{i=1}^{n}(X_i - Y_i)}$

4.2. Manhattan distance was computed by the following formula:

$d= \sum_{i=1}^{n} |X_i - Y_i|$

4.3. Pearson Correlation was computed by the following formula:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

Step 5: The data points with minimum distance from the cluster centers were assigned to the respective cluster.

Step 6: There were two split techniques used in this algorithm are split-simple and split-variance.

6.1. if (split-method = split-simple)

6.2. initialize size, partition = 0;

6.3. for i = 0; i < PL; i++
if (size < partitions[i].getRecords().size())
size = partitions[i].getRecords().size();
partition = i;

6.4. for i = 0; i <SL.size() / 2; i++
Record record = splitRecords.remove(0);
The record is added in the empty partition.

6.5. if (splitMethod == Split-Variance)
currentVariance = 0, v = 0, partition = 0;

6.6. for i = 0; i < partitions.length; i++
v = calculated Variance (partitions[i]); // Variance was calculated based on the i^{th} partition.
if (currentVariance < v)
partition = i;
currentVariance = v;
return partition;

Step 7: Mean were calculated as follows:
for i = 0; i < values.length; i++
means[i] += values[i];
for i = 0; i < means.length; i++
means[i] = means[i]/ records.size();
return means;

Step 8: Variance was calculated as:
for int i = 0; i < values.length; i++
variance [i] += (mean[i] – values[i]) * (mean[i] – values[i]);
sum = 0;
for int i = 0; i < variance.length; i++
variance[i] = variance[i]/ records.size();
sum += variance[i];
return sum;

Step 9: New cluster centers were computed by the following formula:

$CC_i = (1/C_i) \sum_{j=1}^{C_i} X_i$ where X is the selected attribute, C_i is the number of record in the i^{th} epoch and CC is the cluster center in the i^{th} epoch.

Step 9: The distance was recalculated to obtain new cluster center.

Step 10: Two thresholds named constant epochs and same centroids were used; constant epoch was used to determine the stopping condition and the process was stopped when the means did not change.

Output: Final cluster solution.

Table 2 Selected values from the whole dataset (1–30)

S. no.	SCN	CT	UCS	UCSh	MA	SECS	BN	BC	NN	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
4	1016277	6	8	8	1	3	4	3	7	1	2
5	1017023	4	1	1	3	2	1	3	1	1	2
6	1017122	8	10	10	8	7	10	9	7	1	4
7	1018099	1	1	1	1	2	10	3	1	1	2
8	1018561	2	1	2	1	2	1	3	1	1	2
9	1033078	2	1	1	1	2	1	1	1	5	2
10	1033078	4	2	1	1	2	1	2	1	1	2
11	1035283	1	1	1	1	1	1	3	1	1	2
12	1036172	2	1	1	1	2	1	2	1	1	2
13	1041801	5	3	3	3	2	3	4	4	1	4
14	1043999	1	1	1	1	2	3	3	1	1	2
15	1044572	8	7	5	10	7	9	5	5	4	4
16	1047630	7	4	6	4	6	1	4	3	1	4
17	1048672	4	1	1	1	2	1	2	1	1	2
18	1049815	4	1	1	1	2	1	3	1	1	2
19	1050670	10	7	7	6	4	10	4	1	2	4
20	1050718	6	1	1	1	2	1	3	1	1	2
21	1054590	7	3	2	10	5	10	5	4	4	4
22	1054593	10	5	5	3	6	7	7	10	1	4
23	1056784	3	1	1	1	2	1	2	1	1	2
24	1059552	1	1	1	1	2	1	3	1	1	2
25	1065726	5	2	3	4	2	7	3	6	1	4
26	1066373	3	2	1	1	1	1	2	1	1	2
27	1066979	5	1	1	1	2	1	2	1	1	2
28	1067444	2	1	1	1	2	1	2	1	1	2
29	1070935	1	1	3	1	2	1	1	1	1	2
30	1057013	8	4	5	1	2	?	7	3	1	4

Missing attribute value is denoted by "?"

Table 2 shows the data from the BCW dataset (rows 1–30). Of the 30 records, one incomplete record (row 30) was excluded from the calculation. Therefore, a total of 29 records were analyzed for final calculations of foggy k-means. The centroids k random data points were selected as the initial centroid. Two clusters were found ($k = 2$): one for malignant and another for benign. Initially, there was a need of two centroids. In the first step, the application size was 3 as the dataset is partitioned in 10 divisions. In the first iteration, the first three SCN are evaluated. Therefore, the first three records (1000025, 1002945, and 1015425) were not considered for centroid calculation. Thus, the counting was started from SCN = 1016277. So the total record is 27 and after removing one duplicate entry, the final complete record is 26. Based on the above algorithm, the calculated random numbers are as under:

$P1 = 0.91 \quad Y1 = 23$

$P2 = 0.49 \quad Y2 = 13$

Based on the above $Y1$ and $Y2$, rows 23 (S. no. 26) and 13 (S. no. 16) were selected for the $K1$ and $K2$. Attribute 2 means second column UCS was selected for this experiment. The first centroid obtained was 2, 4.

The centroid was then updated according to the cluster. The second attribute value was added to update the centroid as shown below in Table 3. SCN means the sample code number rows from the above table, before centroid means the previous centroid obtained, attribute shows the value of UCS attribute selected in this process, and after centroid is the addition of before centroid and the second attribute. Based on the after centroid, CS1 is calculated.

The process was continued till epoch-1 was achieved, and the positive predictive value (PPV) was calculated (Table 4).

The same 29 records were used for the calculation of random k-means (Table 2). To initialize the centroids, k random data points were selected as the initial centroid. In this approach, two clusters ($k = 2$; malignant and benign)

Table 3 Initial centroid 1 calculation based on the first epoch (foggy centroid calculation mechanism)

S. no.	SCN	Before centroid	Attribute	After centroid
1	1016277, 6, 8, 8, 1, 3, 4, 3, 7, 1, 2	2	8	10
2	1018099, 1, 1, 1, 1, 2, 10, 3, 1, 1, 2	10	1	11
3	1018561, 2, 1, 2, 1, 2, 1, 3, 1, 1, 2	11	1	12
4	1035283, 1, 1, 1, 1, 1, 1, 3, 1, 1, 2	12	1	13
5	1036172, 2, 1, 1, 1, 2, 1, 2, 1, 1, 2	13	1	14
6	1044572, 8, 7, 5, 10, 7, 9, 5, 5, 4, 4	14	7	21
7	1047630, 7, 4, 6, 4, 6, 1, 4, 3, 1, 4	21	4	25
8	1050670, 10, 7, 7, 6, 4, 10, 4, 1, 2, 4	25	7	32
9	1050718, 6, 1, 1, 1, 2, 1, 3, 1, 1, 2	32	1	33
10	1056784, 3, 1, 1, 1, 2, 1, 2, 1, 1, 2	33	1	34
11	1057013, 8, 4, 5, 1, 2, 1, 7, 3, 1, 4	34	4	38
12	1066373, 3, 2, 1, 1, 1, 1, 2, 1, 1, 2	38	2	40
13	1066979, 5, 1, 1, 1, 2, 1, 2, 1, 1, 2	40	1	41

$CS1 = (1/C_i)\sum_{j=1}^{C_i} X_i = 41/13 = 3.1538$

Table 4 PPV result based on foggy centroid

*Y*1	*P*1	*Y*2	*P*2	*K*1	*K*2	CS1	CS2	CL1	CL2	PPV
23	0.91	13	0.49	2	4	3.15	2.64	3.19	2.58	1.00
22	0.87	10	0.39	2	1	2.61	2.00	2.62	2.11	0.66
20	0.79	7	0.29	4	1	3.30	2.64	3.21	2.80	1.00
20	0.80	12	0.46	4	4	3.30	2.78	3.21	2.73	1.00
25	0.96	19	0.74	1	1	2.61	2.50	2.74	2.62	1.00
14	0.54	17	0.65	7	3	3.00	3.07	2.69	3.05	1.00
19	0.74	15	0.60	1	4	2.30	3.00	2.45	2.89	1.00
7	0.28	15	0.61	1	4	2.23	3.14	2.38	3.02	1.00
10	0.40	13	0.50	1	1	2.38	3.14	2.55	3.25	0.66
22	0.87	5	0.21	1	10	2.38	3.92	2.56	3.40	1.00

Total number of records = 30
Final records = 29 (after incomplete removal)
Parameters: Euclidean, split-simple, epoch = 4, attribute used = 2, number of iteration = 10

were obtained that needed two centroids initially. In the first step, the application size was 3. In the first iteration, the first three SCN are evaluated. Therefore, first three records 1000025, 1002945, and 1015425 were not considered for the centroid calculation. Thus, the counting was started from SCN = 1016277. In random approach, random numbers were not required since initial centroid was 0, 0 but the row selection is random. Then, the next centroid was calculated according to the cluster row values as shown in Table 5. Based on the after centroid CS1 is calculated.

The process was continued till epoch-1, and the PPV was calculated (Table 6).

Table 5 Initial centroid 1 calculation based on the first epoch (random centroid calculation mechanism)

S. no.	Before centroid	Attribute	After centroid
1	0	3	3
2	3	1	4
3	4	7	11
4	11	4	15
5	15	1	16
6	16	1	17
7	17	7	24
8	24	1	25
9	25	3	28
10	28	5	33
11	33	1	34
12	34	4	38
13	38	1	39
14	39	2	41
15	41	2	43
16	43	1	44
17	44	1	45
18	45	1	46

$CS1 = (1/C_i)\sum_{j=1}^{C_i} X_i = 46/18 = 2.5555$

Results

In this section, the results are discussed for different cases considering variable parameters. A total 699 instances of BCW dataset, each having nine input attributes (2–10) and one target attribute (11) which is either benign or malignant, have been considered for the experimentation. There are 16 incomplete and 8 duplicate records in total instance. So after removal of these 24 attributes, 675 records are left for the experimentation.

In case 1, the results were obtained using centroid (foggy/random), distance (Euclidean), split (simple), threshold (constant epoch), epoch (4, 5, 6, 7, 9), BCW attribute

Table 6 PPV result based on random centroid

*K*1	*K*2	CS1	CS2	CL1	CL2	PPV
2.55	2.88	2.55	2.88	3.19	2.58	1.00
2.55	1.44	2.55	1.44	2.62	2.11	0.66
2.55	3.22	2.55	3.22	3.21	2.80	1.00
2.55	3.11	2.55	3.11	3.21	2.72	1.00
2.16	3.11	2.16	3.11	2.74	2.62	1.00
2.44	3.11	2.44	3.11	2.69	3.05	1.00
2.16	3.11	2.16	3.11	2.45	2.89	1.00
2.22	3.11	2.22	3.11	2.38	3.02	1.00
2.50	3.11	2.50	3.11	2.55	3.25	0.66
2.61	3.11	2.61	3.11	2.56	3.39	1.00

Total number of records = 30
Final records = 29 (after incomplete removal)
Parameters: Euclidean, split-simple, epoch = 4, attribute used = 2, number of iteration = 10

Table 7 Case 1 (based on epoch)

S. no.	Attributes	Measures/values
1	Centroid	Foggy (F), random (R)
2	Distance	Euclidean
3	Split	Simple
4	Threshold	Constant epoch
5	Epoch (E)	4, 5, 6, 7, 9 (3–10)
6	BCW attribute (A)	2 (1–9)
7	Number of iteration (I)	10 (4–0)

(2), and iteration (10) properties. In foggy centroid, the first centroid was calculated based on random values and for random centroid, the initial centroid was considered as (0, 0). The Euclidean distance algorithm was used to compare the distance between cluster centers and data point. The divisive clustering techniques to split clusters, although received less attention, might have a remarkable impact on overall clustering results. In this approach, the cluster with more elements was split to yield good clustering. The cluster with the highest centroid variance was considered in the variance split. Constant epoch is considered if the epoch values are decided by us. In this case, the epoch variations were 4, 5, 6, 7, and 9. The constant epoch determined the stopping condition. In case of same centroid, the process was stopped when means remained same. The positive predictive value (PPV) was calculated as mentioned below and compared between groups.

$$\text{PPV} = \frac{\text{number of true positive}}{\text{number of true positive} + \text{number of false positive}}$$

where a "true positive" indicates positive prediction from test results and positive result for that case, while a "false positive" indicates a positive prediction and the related case has a negative result. The results for case 1 are presented in Table 7, and its PPV is presented in Table 8. The variations observed in total PPV were due to random initialization of the centroid and not due to epoch variations. Since no variations were observed in the random centroid selection, initialization remained same. The results presented in the half part (column 6–10) of the Table 8 (R, E-4; R, E-5; R, E-6; R, E-7; R, E-9) indicate that the epoch variations did not affect the PPV. The total value indicates the total PPV from the 10 iterations.

In case 2, the results were obtained based on the centroid (foggy/random), distance (Euclidean), split (simple), threshold (constant epoch), epoch (4), BCW attribute (1, 3, 4, 5, 6, 7, 8, 9), and iteration (10). These results are presented in Table 9, and PPVs for case 2 are shown in Tables 10 and 11. The variation observed in total PPV was due to random initialization and attribute value variations and possibly due to variations observed in random centroid. This suggests that the attribute variations do affect PPV. This method returns, for each cluster, a centroid one for each attribute, and a rule describing the position of the data in the cluster. It denotes the mode for the mean and variance for the numerical attributes.

Table 8 Effect of k-means considering variable epoch with foggy and random centroid

F, E-4	F, E-5	F, E-6	F, E-7	F, E-9	R, E-4	R, E-5	R, E-6	R, E-7	R, E-9
0.73	0.65	0.73	0.73	0.65	0.73	0.73	0.73	0.73	0.73
0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
0.84	0.95	0.95	0.95	0.88	0.95	0.95	0.95	0.95	0.95
0.92	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
0.69	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.97	0.97	0.97
0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Total = 8.86	8.85	8.94	8.95	8.79	8.95	8.95	8.95	8.95	8.95

Foggy centroid (F), Epoch value is 4 (E-4), Random centroid (R)

Table 9 Case 2 (based on attributes)

S. no.	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Euclidean
3	Split	Simple
4	Threshold	Constant epoch
5	Epoch (E)	4 (3–10)
6	Attribute (A)	1, 3, 4, 5, 6, 7, 8, 9 (1–9)
7	Number of iteration (I)	10 (4–10)

Therefore, the attribute values determine the centroids for each iteration and also affect the PPV.

In case 3, the results were obtained based on the centroid (foggy/random), distance (Euclidean), split (simple), threshold (constant epoch), epoch (4), BCW attribute (2, 3, 4, 5), and iteration (4, 5, 6, 7, 8). The results for case 3 are shown in Table 12, and PPVs are shown in Tables 13, 14, 15, and 16. The iterations did not cause variations, but determined the partitions. The partitions were increased with increase in iterations. Though the number of iterations was higher, due to less attributes in each partition, the attributes were less which reduced the computation time.

The results for case 4 were obtained based on the centroid (foggy/random), distance (Euclidean), split (simple), BCW attributes (2, 3, 4, 5, 6, 7, 8, and 9), and iteration (10). The results for case 4 are presented in Table 17, and PPV calculated based on it is presented in Table 18. The process was stopped when means remained same for the same centroid. The process was not restricted to the epoch, and the possibility of better cluster selection was improved as suggested by significantly improved PPV.

The results for case 5 were obtained based on the centroid (foggy/random), distance (Euclidean), split (variance), BCW attribute (2, 3, 4, 5, 6, 7, 8, and 9), and iteration (10). The results for case 5 are presented in Table 19, and PPV calculated based on that is shown in Table 20. The highest

Table 10 Effect of k-means considering variable attributes with foggy centroid

F, E-4, A-1	F, E-4, A-3	F, E-4, A-4	F, E-4, A-5	F, E-4, A-6	F, E-4, A-7	F, E-4, A-8	F, E-4, A-9
0.86	0.79	0.60	0.68	0.73	0.72	0.68	0.57
0.84	0.86	0.79	0.86	0.88	0.76	0.85	0.73
0.82	0.95	0.81	0.91	0.91	0.85	0.84	0.69
0.84	0.81	0.73	0.71	0.85	0.82	0.78	0.60
0.85	0.78	0.76	0.65	0.91	0.75	0.75	0.60
0.85	0.94	0.89	0.88	0.95	0.92	0.88	0.82
0.72	0.92	0.89	0.85	0.88	0.95	0.89	0.75
0.89	0.94	0.91	0.85	0.95	0.97	0.91	0.84
0.82	0.92	0.91	0.86	0.91	0.94	0.86	0.69
0.82	0.95	0.94	0.92	0.91	0.98	0.94	0.84
Total = 8.36	8.91	8.28	8.21	8.89	8.71	8.42	7.18

Foggy centroid (F), Epoch value (E), Attribute value (A)

Table 11 Effect of k-means considering variable attributes with random centroid

R, E-4, A-1	R, E-4, A-3	R, E-4, A-4	R, E-4, A-5	R, E-4, A-6	R, E-4, A-7	R, E-4, A-8	R, E-4, A-9
0.86	0.79	0.60	0.72	0.73	0.72	0.72	0.53
0.81	0.86	0.79	0.86	0.88	0.76	0.85	0.72
0.88	0.95	0.81	0.91	0.91	0.85	0.84	0.69
0.84	0.85	0.73	0.71	0.85	0.82	0.78	0.57
0.85	0.78	0.81	0.65	0.91	0.75	0.75	0.56
0.85	0.94	0.89	0.88	0.95	0.94	0.88	0.82
0.72	0.92	0.89	0.86	0.88	095	0.91	0.75
0.82	0.94	0.91	0.85	0.95	0.97	0.91	0.84
0.69	0.92	0.91	0.86	0.91	0.94	0.86	0.69
0.82	0.95	0.94	0.92	0.91	0.98	0.94	0.84
Total = 8.18	8.95	8.33	8.27	8.89	8.72	8.47	7.05

Random centroid (R), Epoch value (E), Attribute value (A)

Table 12 Case 3 (based on iterations)

S. no.	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Euclidean
3	Split	Simple
4	Threshold	Constant epoch
5	Epoch (E)	4 (3–10)
6	Attribute (A)	2, 3, 4, 5 (1–9)
7	Number of iteration (I)	4, 5, 6, 7, 8 (4–10)

variance with respect to its centroid was considered for the variance split. The combination of highest variance and same centroid provided significantly improved PPV.

The results for case 6 were obtained based on the centroid (foggy/random), distance (Manhattan, Pearson coefficient), split (simple), threshold (constant epoch), epoch (4), BCW attribute (2, 3, 4, 5, 6, 7, 8, 9), and iteration (10). The results for case 6 are presented in Table 21, and PPV is shown in Tables 22 and 23. Distance algorithm was used to calculate the similarity between the different data points obtained based on distance metric. In this

Table 13 Effect of k-means considering variable iterations with foggy centroid and A-2

F, E-4, A-2, I-4	F, E-4, A-2, I-5	F, E-4, A-2, I-6	F, E-4, A-2, I-7	F, E-4, A-2, I-8
0.65	0.73	0.65	0.73	0.65
0.85	0.85	0.78	0.85	0.85
0.95	0.95	0.95	0.95	0.95
0.76	0.84	0.76	0.84	0.84
NA	0.69	0.69	0.75	0.75
NA	NA	0.97	0.97	0.97
NA	NA	NA	0.95	0.95
NA	NA	NA	NA	0.95
Total = 3.23	4.08	4.82	6.07	6.94

Foggy centroid (F), Epoch value (E), Attribute value (A), Iteration value (I)

Table 14 Effect of k-means considering variable iterations with foggy centroid and A-3

F, E-4, A-3, I-4	F, E-4, A-3, I-5	F, E-4, A-3, I-6	F, E-4, A-3, I-7	F, E-4, A-3, I-8
0.79	0.79	0.79	0.79	0.79
0.86	0.86	0.86	0.86	0.79
0.95	0.95	0.95	0.95	0.89
0.85	0.85	0.85	0.85	0.85
NA	0.73	0.78	0.73	0.78
NA	NA	0.94	0.94	0.94
NA	NA	NA	0.92	0.89
NA	NA	NA	NA	0.92
Total = 3.47	4.21	5.20	6.08	6.89

Foggy centroid (F), Epoch value (E), Attribute value (A), Iteration value (I)

Table 15 Effect of k-means considering variable iterations with foggy centroid and A-4

F, E-4, A-4, I-4	F, E-4, A-4, I-5	F, E-4, A-4, I-6	F, E-4, A-4, I-7	F, E-4, A-4, I-8
0.60	0.60	0.60	0.60	0.60
0.79	0.79	0.79	0.79	0.79
0.81	0.81	0.81	0.81	0.81
0.69	0.73	0.73	0.69	0.73
NA	0.81	0.81	0.81	0.81
NA	NA	0.89	0.88	0.89
NA	NA	NA	0.89	0.89
NA	NA	NA	NA	0.91
Total = 2.91	3.76	4.66	5.50	6.47

Foggy centroid (F), Epoch value (E), Attribute value (A), Iteration value (I)

Table 16 Effect of k-means considering variable iterations with random centroid and A-2

R, E-4, A-2, I-4	R, E-4, A-2, I-5	R, E-4, A-2, I-6	R, E-4, A-2, I-7	R, E-4, A-2, I-8
0.73	0.73	0.73	0.73	0.73
0.85	0.85	0.85	0.85	0.85
0.95	0.95	0.95	0.95	0.95
0.84	0.84	0.84	0.84	0.84
NA	0.75	0.75	0.75	0.75
NA	NA	0.97	0.97	0.97
NA	NA	NA	0.95	0.95
NA	NA	NA	NA	0.95
Total = 3.39	4.14	5.11	6.07	7.02

Random centroid (R), Epoch value (E), Attribute value (A), Iteration value (I)

Table 17 Case 4 (based on same centroid)

S. no.	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Euclidean
3	Split	Simple
4	Threshold	Same centroid
5	Attribute (A)	2, 3, 4, 5, 6, 7, 8, 9 (1–9)
6	Number of iteration (I)	10

Table 18 Effect of k-means considering split-simple with foggy and random centroid

A-2, I-10	A-3, I-10	A-4, I-10	A-5, I-10	A-6, I-10	A-7, I-10	A-8, I-10	A-9, I-10
Foggy centroid							
0.67	0.85	0.71	0.70	0.77	0.65	0.83	0.67
0.88	0.92	0.88	0.88	0.92	0.74	0.86	0.82
0.95	0.89	0.92	0.94	0.91	0.77	0.85	0.77
0.74	0.80	0.79	0.76	0.85	0.73	0.79	0.62
0.86	0.79	0.85	0.67	0.91	0.71	0.71	0.68
0.95	0.92	0.89	0.88	0.95	0.83	0.91	0.88
0.92	0.92	0.71	0.91	0.88	0.94	0.91	0.79
0.98	0.98	0.95	0.92	0.95	0.97	0.92	0.88
0.92	0.95	0.92	0.95	0.92	0.97	0.86	0.77
1.00	0.94	0.98	0.97	0.92	0.95	0.95	0.92
Total = 8.91	8.99	8.64	8.59	9.01	8.29	8.62	7.83
Random centroid							
0.76	0.80	0.71	0.74	0.77	0.74	0.76	0.62
0.85	0.86	0.85	0.86	0.89	0.74	0.83	0.80
0.95	0.95	0.88	0.91	0.94	0.85	0.83	0.77
0.82	0.83	0.79	0.70	0.88	0.83	0.79	0.62
0.77	0.79	0.85	0.67	0.88	0.73	0.74	0.65
0.97	0.95	0.91	0.89	0.95	0.95	0.88	0.89
0.94	0.91	0.85	0.85	0.88	0.94	0.92	0.76
0.97	0.95	0.95	0.86	0.97	0.98	0.92	0.85
0.92	0.92	0.92	0.86	0.92	0.94	0.86	0.71
1.00	0.95	0.98	0.94	0.92	0.98	0.94	0.92
Total = 8.96	8.95	8.71	8.31	9.03	8.71	8.50	7.64

Attribute value (A), Iteration value (I)

Table 19 Case 5 (based on variance)

S. no.	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Euclidean
3	Split	Variance
4	Threshold	Same centroid
5	Attribute (A)	2, 3, 4, 5, 6, 7, 8, 9 (1–9)
6	Number of iteration (I)	10

Table 20 Effect of k-means considering split-variance with foggy and random centroid

A-2, I-10	A-3, I-10	A-4, I-10	A-5, I-10	A-6, I-10	A-7, I-10	A-8, I-10	A-9, I-10
Foggy centroid							
0.67	0.80	0.77	0.82	0.73	0.71	0.89	0.67
0.88	0.92	0.88	0.88	0.88	0.79	0.86	0.82
0.95	0.95	0.92	0.94	0.94	0.79	0.86	0.68
0.82	0.91	0.82	0.70	0.92	0.86	0.79	0.71
0.86	0.85	0.79	0.76	0.91	0.73	0.85	0.68
0.95	0.92	0.91	0.88	0.95	0.97	0.88	0.88
0.95	0.92	0.88	0.91	0.89	0.94	0.91	0.74
0.95	0.98	0.97	0.92	0.95	0.97	0.92	0.85
0.97	0.95	0.83	0.95	0.91	0.97	0.88	0.68
0.98	0.95	0.97	0.97	0.91	0.95	0.95	0.89
Total = 9.01	9.19	8.76	8.74	9.01	8.70	8.82	7.64
Random centroid							
0.67	0.70	0.64	0.70	0.73	0.65	0.71	0.55
0.77	0.79	0.77	0.79	0.88	0.71	0.79	0.70
0.88	0.89	0.80	0.83	0.91	0.79	0.82	0.68
0.76	0.80	0.70	0.64	0.85	0.73	0.79	0.56
0.70	0.73	0.77	0.62	0.91	0.70	0.70	0.55
0.95	0.92	0.88	0.89	0.95	0.94	0.86	0.82
0.92	0.91	0.86	0.83	0.88	0.91	0.89	0.74
0.95	0.91	0.92	0.86	0.95	0.97	0.91	0.85
0.89	0.91	0.85	0.79	0.91	0.94	0.82	0.70
1.00	0.97	0.94	0.88	0.91	0.95	0.94	0.85
Total = 8.52	8.55	8.16	7.86	8.89	8.31	8.25	7.02

Attribute value (A), Iteration value (I)

Table 21 Case 6 (based on distance metrics)

S. no.	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Manhattan, Pearson coefficient
3	Split	Simple
4	Threshold	Same centroid
5	Epoch (E)	4 (3–10)
6	Attribute (A)	2, 3, 4, 5, 6, 7, 8, 9 (1–9)
7	Number of iteration (I)	10

study, Euclidean, Manhattan, and Pearson coefficients were used and less differences in PPV were obtained with only Euclidean and Manhattan. Thus, the clustering algorithm combined with these distance metrics found to be beneficial. However, the combination of k-means with Pearson correlation led to non-intuitive centroid that represented

Table 22 Effect of k-means based on epoch and distance variations

E-4, A-2, I-10	E-4, A-3, I-10	E-4, A-4, I-10	E-4, A-5, I-10	E-4, A-6, I-10	E-4, A-7, I-10	Case
Total = 8.77	8.92	8.22	8.29	8.89	8.50	Epoch and Manhattan distance with foggy centroid
Total = 6.56	7.62	7.71	7.20	7.49	7.04	Epoch and Pearson coefficient with foggy centroid
Total = 8.96	8.95	8.37	8.29	8.89	8.70	Epoch and Manhattan distance with random centroid
Total = 7.88	7.67	7.65	7.31	7.37	7.22	Epoch and Pearson coefficient with random centroid

Epoch (E), Attribute value (A), Iteration value (I)

Table 23 Effect of k-means based on same centroid and distance variations

A-2, I-10	A-3, I-10	A-4, I-10	A-5, I-10	A-6, I-10	A-7, I-10	Case
Total = 9.08	9.06	8.67	8.50	8.93	8.71	Same centroid and Manhattan distance with foggy centroid
Total = 7.71	7.98	7.56	7.20	7.67	6.89	Same centroid and Pearson coefficient with foggy centroid
Total = 9.07	9.12	8.70	8.29	9.03	8.70	Same centroid and Manhattan distance with random centroid
Total = 7.88	7.67	7.65	7.31	7.37	7.22	Same centroid and Pearson coefficient with random centroid

Attribute value (A), Iteration value (I)

Table 24 Case 7 (based on distance metrics, variance, and same centroid)

S.no	Attributes	Measures
1	Centroid	Foggy (F), random (R)
2	Distance	Manhattan, Pearson coefficient
3	Split	Variance
4	Threshold	Same centroid
5	Attribute (A)	2, 3, 4, 5, 6, 7 (1–9)
6	Number of iteration (I)	10 (4–10)

Table 25 Effect of k-means considering based on same centroid, same variance, and distance variations

A-2, I-10	A-3, I-10	A-4, I-10	A-5, I-10	A-6, I-10	A-7, I-10	Case
Total = 9.05	9.10	8.49	8.85	9.02	8.62	Same centroid, same variance and Manhattan distance with foggy centroid
Total = 7.35	7.71	7.44	7.07	7.58	7.40	Same centroid, same variance and Pearson coefficient with foggy centroid
Total = 8.52	8.55	8.16	7.86	8.89	8.31	Same centroid, same variance and Manhattan distance with random centroid
Total = 7.95	7.43	7.62	7.59	8.38	7.22	Same centroid, same variance and Pearson coefficient with random centroid

Attribute value (A), Iteration value (I)

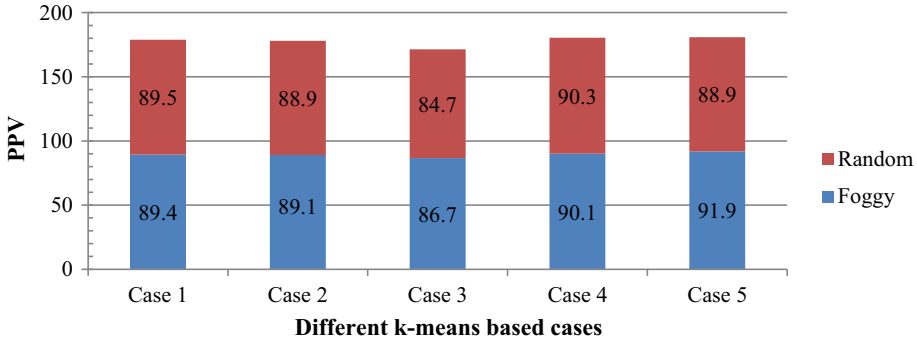

Fig. 1 Case comparison with different attributes

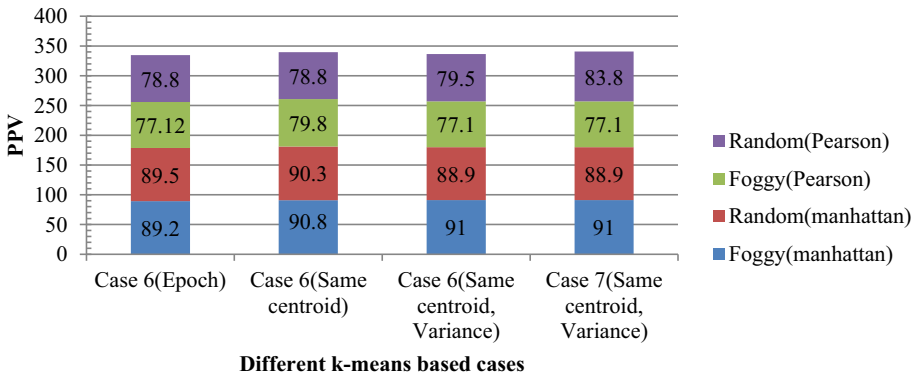

Fig. 2 Case comparisons with different distance algorithms

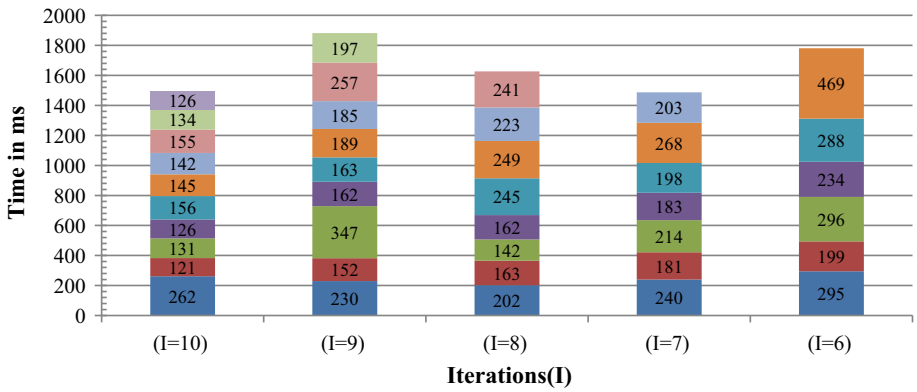

Fig. 3 Time comparison with variable partitions ($I = 10$ to $I = 6$)

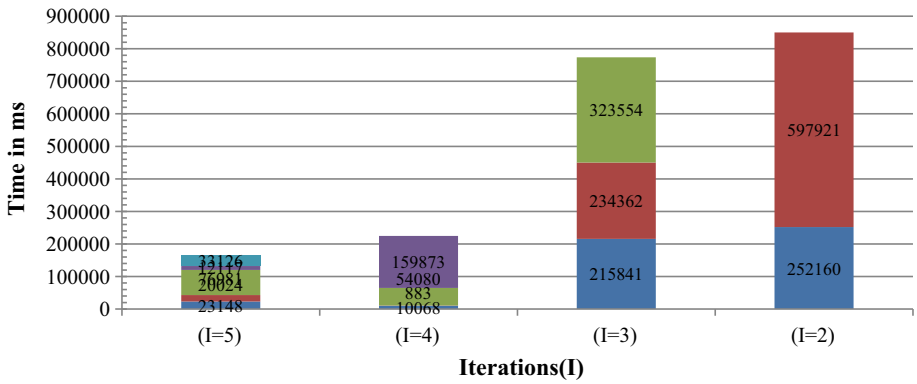

Fig. 4 Time comparison with variable partitions ($I = 5$ to $I = 2$)

the mean of the cluster and Pearson, i.e., anti-correlated objects.

The results for case 7 were obtained based on the centroid (foggy/random), distance (Manhattan, Pearson coefficient), split (variance), threshold (same centroid), BCW attributes (2, 3, 4, 5, 6, and 7), and iteration (10). The results for case 7 are presented in Table 24, and based on this, the PPV was calculated (Table 25). It was based on distance algorithm,

split-variance, and the same centroid. The combination of the highest variance and same centroid can produce significantly improved PPV.

The overall results are summarized in Figs. 1 and 2. The data in Figs. 1 and 2 show that there was approximately 92 % positive prediction accuracy obtained with this approach. The PPV was found to be significantly improved when same centroid and highest variance were combined. In addition, the distance measure standards influenced the PPV values (Fig. 2). The Euclidean and Manhattan were found to be better than Pearson correlation. The PPV computation time is shown in each iteration from iteration $(I) = 2$ to 10 (Figs. 3, 4). The numbers of iterations increase the partitions, and higher number of partitions contains less attributes in comparison with less number of partitions, thereby reducing the computation time (Figs. 3, 4).

Discussion

In this study, the effects of various parameters on k-means clustering algorithm with foggy and random centroid were investigated and the key findings are as follows:

1. The results of the random initialization of the centroid were responsible for the variation in the total PPV and not the epoch variations.
2. The attribute values of BCW dataset determined the mode for the mean and variance of centroid which affected the PPV.
3. Increase in number of iterations increased partitions which contain less attributes, thereby reducing the computation time.
4. For the same centroid, the process was stopped when means remained constant. Thus, the process was not restricted to the epoch and the possibility of better cluster selection was improved.
5. The combination of the highest variance and the same centroid significantly improved the PPV.
6. The combination of Euclidean or Manhattan with k-means algorithm reduced differences in PPV. However, the results obtained by the combination of k-means with the Pearson coefficient were less significant than with Euclidean or Manhattan coefficient.
7. The consistency and uniformity of the results after several repetitions suggest that this clustering approach is relatively efficient.

Replications and future directions

This experiment was performed by using Java version 6 with the help of NetBeans IDE 7.2. The entire experimentation framework was designed and developed by the help of NetBeans IDE. This experiment was performed using the BCW diagnostic dataset. The experimental results were obtained using six attributes, namely centroid (foggy/random), distance algorithms, split methods, thresholds, BCW attributes, and iteration. These experiments can be replicated and controlled based on the attribute values. The replications can be done with the help of attributes and tuned with different attribute values, which are the control parameters of the experimentation. This experiment can be replicated and enhanced by changing in centroid calculation like Pillar algorithm for centroid initialization and Minkowski and Cosine distance algorithms. It can be compared with other clustering algorithms like fuzzy c-means clustering and hierarchical clustering using the same attributes and control parameters to verify better clustering algorithm for BCW dataset.

Conclusions

K-means clustering algorithm, used in this study for classification of BCW dataset, is an unsupervised learning algorithm. It is a simple and easy approach to classify dataset using k clusters. In this study, a computational formulation of integrative clustering using multi-variant parameters was investigated which has accurately classified a BCW dataset. The results suggest that the Euclidean/Manhattan distance algorithm with largest variance and same centroid is a better choice for accurate classification of the dataset. This implies that k-means with the variations applied in our approach is capable to classify the BCW dataset with single and multiple iterations.

Compliance with ethical standards

Conflict of interest The authors Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain declare that they have no conflict of interest. The manuscript does not contain clinical studies or patient data.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136(5):E359–E386
2. Dubey AK, Gupta U, Jain S (2015) Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pac J Cancer Prev 16(10):4237–4245
3. Dubey AK, Gupta U, Jain S (2014) A Survey on Breast Cancer Scenario and Prediction Strategy. In: Proceedings of the 3rd international conference on frontiers of intelligent computing: theory and applications (FICTA), 2014. Springer International Publishing, pp 367–375
4. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of 20th international conference on very large data bases, VLDB 1994. vol 1215. pp 487–499

5. Jain R (2015) Introduction to data mining techniques. http://www.iasri.res.in/ebook/expertsystem/datamining.pdf. Accessed 22 April 2015
6. Alpaydin E (2014) Introduction to machine learning. MIT press, Cambridge, Massachusetts, United States
7. Bradley PS, Fayyad UM (1998) Refining initial points for k-means clustering. In: Proceedings of the 15th international conference on machine learning (ICML), Morgan Kaufmann, San Francisco, vol 98. pp 91–99
8. Mary C, Raja SK (2009) Refinement of clusters from k-means with ant colony optimization. J Theor Appl Inf Technol 6(4):28–32
9. Wang C, Machiraju R, Huang K (2014) Breast cancer patient stratification using a molecular regularized consensus clustering method. Methods 67(3):304–312
10. Rahideh A, Shaheed MH (2011) Cancer classification using clustering based gene selection and artificial neural networks. In: IEEE 2nd international conference on control, instrumentation and automation (ICCIA), 2011. pp 1175–1180
11. Vanisri D, Loganathan C (2010) Fuzzy pattern cluster scheme for breast cancer datasets. In: IEEE international conference on communication and computational intelligence (INCOCCI), 2010. pp 410–414
12. Festa P (2013) A biased random-key genetic algorithm for data clustering. Math Biosci 245(1):76–85
13. Chen CH (2014) A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Appl Soft Comput 20:4–14
14. Wei D, Jiang Q, Wei Y, Wang S (2012) A novel hierarchical clustering algorithm for gene sequences. BMC Bioinform 13(1):174
15. Ahmad FK, Yusoff N (2013) Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. In: IEEE 13th international conference on intelligent systems design and applications (ISDA), 2013. pp 121–125
16. Bache K, Lichman M (2013) UCI machine learning repository. 1990:92. http://archive.ics.uci.edu/ml