International Journal of Soft Computing 11 (5): 322-333, 2016
ISSN: 1816-9503
© Medwell Journals, 2016

Diagnosing Breast Cancer Using Clustering with Feature Selection

[1]Israa Abdulqader, [1]Sherihan Abuelenin and [2]Ahmed Aboelfetouh
[1]Department of Computer Science,
[2]Department of Information System, Mansoura University, Mansoura, Egypt

Abstract: Breast cancer is one of the popular cancers in women and is considered one of the popular causes of death. Earlier detection and diagnosis may save lives and make efficient of life. In this study, a new method for breast cancer diagnosis is proposed. The proposed method consists of three stages: the first divides dataset to two clusters using kernel k-means clustering, the second minimizes features by applying feature selection algorithm on each cluster and the third collects resulting feature from each cluster together and measures the quality using different classifiers. The proposed approach is evaluated using datasets for breast cancer: Breast cancer wisconsin diagnostic dataset "WDBC" get from UCI machine learning repository. The performance of the proposed method is evaluated by measuring accuracy, sensitivity, specificity, mean squared error and time. The experiments are done with three classifiers Naive Bayes "NB", Multilayer Perceptron "MLP" and decision tree J48.

Key words: Clustering, feature selection, classification, breast cancer, clustring, algorthim

INTRODUCTION

Machine learning is a computer science subfield which developed from studying pattern recognition and theory of computational learning in AI "Artificial Intelligence". It is the study branch which provides computers with capability for learning with no need to be an explicit program. Machine learning is concerned with the algorithms' study and construction which could learn from and make data predictions. These algorithms' function is done by means of forming a model from example inputs so as to make data-driven estimates or conclusions, before firmly following static program commands.

There is a typical classification of machine learning tasks into three main classes: supervised learning, unsupervised learning and reinforcement learning. Supervised learning where the computer is presented with example inputs and their preferred outputs, specified by a "teacher", the aim is to acquire an overall instruction which charts inputs to outputs. Learning which is unsupervised where there are no labels presented algorithm of learning, leaving it single to find structure in its input. Unsupervised learning itself could be an aim (realizing patterns that hidden in data) or a means to an end "feature learning". Support learning is where there is an interaction between a computer program and an environment which is dynamic where it is essential to achieve a specific goal (like driving a vehicle) with no need for a teacher explicitly telling it whether it is nearby its aim.

Machine learning methods are being used progressively in making applicable estimates and implications on separate subjects neuroimaging scan data. The rest of this section will explain diverse machine learning methods that were utilized for clustering and feature selection to reduce unwanted feature to make the good classification of feature after selection (Osareh and Shadgar, 2010).

Cluster analysis or clustering is the job in which there is a grouping of a set of items in such a way where same group items "known as cluster" are additionally parallel (in a way or another) to one another than to those in further groups (clusters). Its achievement could happen by means of different algorithms which are clearly different in their concept of what creates a cluster and how to allocate them capably (Hamad and Biela, 2008). The suitable clustering algorithm and parameter settings rely on the single data set and envisioned utilization of the results. Cluster analysis by itself is a task which is not automatic, yet an iterative procedure of discovering knowledge or interactive optimization that is multi-objective which includes trial and failure. It is often required for modifying data preprocessing and model parameters till the result attains the anticipated properties (Manikandan *et al*., 2013). Distribution-based clustering, density-based clustering, connectivity based clustering and centroid-based clustering are some of clustering techniques (Mulyono and Ishida, 2014).

Kernel k-means is a centroid-based clustering. Such algorithm makes an application of the same trick as

Corresponding Author: Israa Abdulqader, Department of Computer Science, Mansoura University, Mansoura, Egypt k-means yet with a single variance which is in the distance calculation, the method of kernel is utilized in place of the Euclidean distance. The kernel function's type is achieved by this parameter like dot, radial, multiquadric, polynomial, epachnenikov, gaussian combination, neural and anova (Tzortzis and Likas, 2008).

K-means with euclidean distance expects the data distribution into elliptical regions. When the assumption doesn't hold, some kind of transformations may be implemented to the data, mapping them to a new space where the learning machine can be used. Kernel function provides a means to define the transformation (Ferreira and Carvalho, 2014).

Suppose there are set of samples $x_1, x_2, ..., x_N$ where $x_i є R^D$ and a mapping function Π that maps x_i from the input space R_D to a new space Q (Cristianini and Taylor, 2000) and definition of kernel function as the dot product in the new space Q:

$$H(xi, x j) = \phi(xi).\ \phi\ (xj) \tag{1}$$

In kernel function there is an important fact that it is constructed without knowing namely the concrete form of Π, implicitly the transformation is defined. Kernel functions are used commonly as shown below (Liu *et al.*, 2012). Radial:

$$H(x_i, x_j) = \exp(-r\|x_i - x_j\|^2) \tag{2}$$

Multiquadric:

$$H(x_i, x_j) = \|x_i - x_j\|^2 + c^2 \tag{3}$$

In classification tasks feature selection plays an important role which a vital reason of using relief-F, CFS and IG. Relief-F is a feature choice procedure that picks instances haphazardly and changed the weights of the feature importance in light of the closest neighbor. By its benefits, Relief-F is a standout amongst the best systems in feature determination (Liu and Motoda, 2008). Its qualities are that it is not reliant on heuristics, keeps running in low-request polynomial time (Kononenko *et al.*, 1997). Furthermore, it is clamor tolerant and powerful to attribute cooperation and being relevant for double or nonstop information; notwithstanding, it doesn't segregate between excess attributes and low amounts of preparing instances bluffing the algorithm (Kira and Rendell, 1992). The function of Relief-F is:

$$\text{Function [out]} = \text{fs Relief F}(X,\ Y,\ k,\ m) \tag{4}$$

Where:

X = Features on current trunk
Y = Label of instances
k = Size of the neighborhood that wish to be evaluated and m is how many samples you want to try

Like the greater part of feature determination programs, CFS (Correlation based Feature Selection) utilizes an inquiry algorithm alongside a function to assess the value of feature subsets. The heuristic by which CFS measures the integrity of feature subsets considers the redundant individual features for foreseeing the label mark alongside the stage of intercorrelation among them. The heuristic theory based can be expressed as (Han *et al.*, 2011).

Great feature subsets include features exceedingly corresponded (prescient of) with the class, yet uncorrelated with (not prescient of) one another. Next mathematical statement formalizes the heuristic:

$$G_s = \frac{K\overline{r_{cf}}}{\sqrt{K + K(K-1)\overline{\overline{r_{ff}}}}} \tag{5}$$

Where:

G_s = The "G" of a feature subset
S = Including K features ¯(r_cf)
¯(r_cf) = The mean feature class correlation (f∈S)
¯(r_ff) = average feature-feature intercorrelation

It is in fact, pearson connection where the sum total of what variables has been institutionalized. The numerator can be considered as giving a sign of how prescient of the class a gathering of features are the denominator of the amount of excess there is among them. The heuristic goodness apportion ought to filter unimportant attributes as they will be poor indicators of the class (Han *et al.*, 2011). Repetitive features ought to be disregarded as they will be exceptionally corresponded with one or a greater amount of alternate features (Hall and Smith, 1997).

All things considered most classification undertakings in ML include learning to recognize ostensible class values, however might include features that are ordinal or persistent. A measure in light of restrictive entropy is utilized to quantify relationships in the middle of features and the class and between features. Ceaseless features are first changed over to ostensible by binning (Press, 1988). Various feature positioning and feature determination procedures have been proposed in the machine learning writing. Entropy is a usually utilized as a part of the information hypothesis measure which describes the immaculateness of a self-assertive accumulation of samples. It is in the basis of the Information Gain attribute evaluation (IG) property positioning techniques. It can be considered entropy measure as a measure of framework's eccentrics. The entropy of Y is:

$$H(Y) = -\sum_{y\in Y} p\,(y)\log_2\,(p(y)) \tag{6}$$

where, p(y) is the negligible likelihood thickness function for the arbitrary variable Y. If the watched estimations of Y in the preparation information set S are divided by estimations of a second feature X and the entropy of Y concerning the allotments incited by X is not exactly the entropy of Y before apportioning, then there is a relationship between features Y and X. At that point the entropy of Y after watching X is:

$$H\,(\frac{Y}{X}) = -\sum_{x\in X} p(x) \sum_{y\in Y} p\left(\frac{y}{x}\right)\log_2\left(p\left(\frac{y}{x}\right)\right) \tag{7}$$

where, p(y|x) is the restrictive likelihood of y given x. Given the entropy as a foundation of pollution in a preparation set S, we can define a measure reflecting extra information about Y gave by X that speaks to the sum by which the entropy of Y declines. This measure is known as IG. It is given by:

$$IG = H(Y) - H\left(\frac{Y}{X}\right) = H(X) - H(\frac{X}{Y}) \tag{8}$$

IG (Information Gain) is a symmetrical measure (refer to mathematical statement previous). The information increased about Y after watching X is equivalent to the information increased about X after watching Y (Novakovic, 2009).

Classification accuracy depends on the feature selection and good choices of the method that classify the feature also affect the accuracy of the proposed system. The machine is given an arrangement of preparing samples (x_i, y_i) where x_i the is this present reality information occasions and the y_i are the marks showing which class the occurrence fits in with. For the two class design acknowledgment issue y_i = +1 or y_i-1. A preparation case (x_i, y_i) is called positive if (y_i) and negative something else (Patil and Sherekar, 2013). A Multilayer Perceptron "MLP" is a feedforward neural network classifier that used widely in machine learning problems, it is consists of an input and an output layers with one or multi hidden layers and number of units for each layer (Vilan *et al.*, 2013). Each unit in every layer attached with a specific weight to each unit in the following universal MLPs layers are approximated any real valued functions (Jamuna *et al.*, 2010). Classification problem in a two classes for a given input:

$$z_j = g\left(\sum\nolimits_{d=1}^{D} w_{jd}\ x_d + w_{j0}\right) \tag{9}$$

$$f = h\left(\sum\nolimits_{J=1}^{J} v_j\ z_j + v_0\right) \tag{10}$$

Where:
z_j j = 1, J = The energizing of the hidden-layer units
w_{jd} = The weights between the input and the hidden layer

Similarly, v_j are weights connecting the hidden layer to the output unit f. The terms w_{j0} and v_0 are the biases for the hidden and output units. h (t). and g (t)are continuous sigmoid function, usually of the form tanh(t) or the logistic function (Chan *et al.*, 2002).

Naive bayes are connected into learning errands where every case x is portrayed by a conjunction of property estimations and where the objective capacity f (x) can tackle any worth from some limited set V. An arrangement of preparing illustrations of the objective capacity is given and another occurrence is displayed, portrayed by the tuple of property estimations <a_1, a_2, a_3, ..., a_n>. The learner is approached to anticipate the objective quality for the new case. The bayesian approach used to group the new case is to allot the most plausible target esteem, given the trait values <a_1, a_2, a_3, ..., a_n> that depict the instance (Aloraini, 2012). The function of NB is:

$$V_{NB} = \text{argmax}\prod\nolimits_{j=1}^{n} P\left(a_i \middle| V_j\right)P\left(V_j\right) \tag{11}$$

Decision tree algorithm is routinely used for characterization issue. In this algorithm, the data set is learnt and showed. Along these lines at whatever point another data thing is given for order it will be gathered as requirements to be figured out from the past dataset. Thusly, the algorithm will in like manner learn and models data taking into account the arrangement data. One of the characteristics of decision tree that is it can function admirably with immense data sets. This is indispensable as significant measure of data. It functions admirably in light of the way that decision tree gives the most lifted recognition execution and can create and interpret show adequately (Navada *et al.*, 2011).

The J48 decision tree classifier takes after the partner direct algorithm. Remembering the objective of developing different things, firstly need to build a decision tree in view of the trademark assessment of open get ready data. Subsequently, at whatever point it encounters course of action of things (get ready set) it segregates the property that isolates the different states most unmistakably. That component which claims the limit let's see more around the data events so it might gather as well as can be expected, be say that is astonishing information get. At present, among the comprehensible evaluation of this component, if have some worth for it there is no a dubious, along these lines, the data cases that got to be inside of its order possess the same nature of target variable, then it completes that branch and apportions to it the target estimation that have gotten (Ying *et al*., 2015).

Literature review: Investigated the materialness of decision trees for identification of high-hazard breast cancer gatherings over the dataset created by department of genetics of faculty of medical sciences of universidad nova de lisboa with 164 controls and 94 cases in "WEKA" machine learning apparatus. To measurably approve the affiliation found, change tests were utilized. They discovered a high-hazard breast cancer gathering made out of 13 cases and just 1 control with a Fisher exact test (for acceptance) estimation of 9.7×10-6 and a p-estimation of 0.017. These outcomes demonstrated that it is conceivable to find measurably significant relationship with breast cancer by inferring a decision tree and chosen the best leaf.

Lavanya and Rani (2011) proposed decision tree that has been successfully connected for feature determination to investigate the performance of classifier. The examination of cart classifier is performed with and without feature choice as far as precision, size of the tree and time to construct a model on different breast cancer datasets. The outcomes demonstrate that a specific feature determination utilizing cart has improved the classification precision of a specific dataset.

Badriyah *et al*. (2012) proposed the decision trees model to tackle specific issues that ordinarily utilize logistic regression as an answer, particularly to anticipate danger of death. From tests the conclusion is that logistic regression and decision trees are both effective means of building models to anticipate danger of mortality. Both strategies give sensible separation however decision trees considering the benefits of human interpretability can be seen as a commendable different option for logistic regression in the zone of data mining.

Zheng *et al*. (2014) presented approach that based on the extracted tumor features to diagnose breast cancer disease by developing a hybrid algorithm of K-means and support vector machine. For recognizing the benign and malignant hidden patterns separately the K-means algorithm is used. Calculation and treatments of the membership of each tumor in these pattern is done as a new feature in the training model. Then to obtain the new classifier that differentiates the new tumors a support vector machine is used. The proposed method accuracy was 97-38% based on 10-fold cross validation when tested on the "WDBC" dataset. From 32 original features six tumor features re extracted in the training phase. The results illustrated time saving in the training stage beside breast cancer disease.

Win *et al*. (2014) described a state of-the-art machine learning based approach called averaged one-dependence estimators with subsumption resolution "AODEsr" to tackle the problem of predicting from DNA microarray gene expression data whether a particular cancer will recur within a specific timeframe which is usually 5 year. To lower the computational complexity, they employ an entropy-based gene selection approach to select relevant prognostic genes that are directly responsible for recurrence prediction. This proposed system has achieved an average accuracy of 98.9% in predicting cancer recurrence over 3 datasets.

Thein and Tun (2015) projected a method for the distinguishing of breast cancer among various classes of breast cancer. Such method relies on the wisconsin diagnostic and prognostic breast cancer and the classification of datasets of various types of breast cancer. The projected system implements the method of island-based training to be more accurate and have less time for training by utilizing and analyzing two different migration topologies.

Montazeri *et al*. (2016) proposed a method of classification that is rule-based with the techniques of machine learning for predicting various forms of surviving Breast cancer. Naive Bayes "NB", Trees Random Forest "TRF", 1-Nearest Neighbor "1NN", AdaBoost "AD", Support Vector Machine "SVM", RBF Network "RBFN" and Multilayer Perceptron "MLP" machine learning techniques with 10-cross fold technique have been utilized with the projected model in order to predict surviving breast cancer. In this study, Trees Random Forest "TRF" methods presented better outcomes when compared to other methods "NB, 1NN, AD, RBFN and MLP". There was correctness rate of 96%. Nevertheless, the method of 1NN machine learning provided performance that is poorly accurate of 91%.

MATERIALS AND METHODS

Proposed method: According to the previous researches, most of the researches have been worked on the dataset as it one cluster. The proposed method tries to find a way to get the best result with minimum time execution without affect the overall results of the whole system.

The previous researches reduce the feature from the data set as it without any preprocessing. In the proposed method, we try to find a way to get the optimum result in

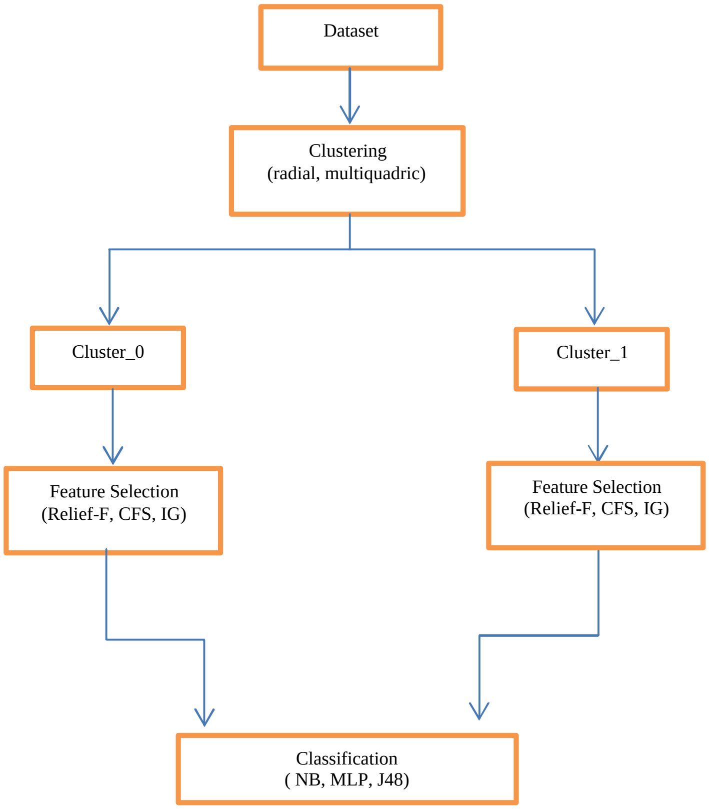

Fig. 1: Proposed method

short time execution. The goals of the proposed method are: reducing training time, reducing classification over fitting and removing the irrelevant features to solve dimensionality problem. The proposed method is shown in Fig. 1.

The proposed framework divides the dataset into two clusters based on kernel K-means using two types of kernel functions; radial and multiquadric. Radial and multi-quadric are used here because of strong tolerance to input noise, ability of online learning and good generalization. The proposed framework works on each cluster separately in the same time which saves time and effort. Each cluster has its own feature which make a value to reduce unimportant or redundant features that not affect the decision of being benign or malignant tumor. Feature selection is a very important step with the system which can be used to make a classification because unimportant feature affects the final result and also wastes time and memory

In the proposed framework, feature selection has been done using relief-F, Correlation Feature Subset selection (CFS) and Information Gain (IG). These methods do not consider the relationships between features in selecting redundant features. For a given feature vector the purpose is to eliminate redundant or irrelevant features. Using feature selection has many advantage like improving understandability and decreasing data cost and handling in machine learning feature selection has great attention because of all these advantage. The reasons of using these techniques are fast, scalable independent of classifier, better and robust to overfitting but it has a drawback that is computational complexity.

Each cluster has its reduced vector of feature. Now the classification will take place but the question here either the classifier works on each selected feature separately or fuses these features. Using the union of the two features, it will get one feature vector. The intersection can't be used because the two feature vector may have not any intersection between each other which make it is illogic. Now get the union of the features after selection then the classifier will take place. By multilayer perceptron, naïve bayes and decision tree j48 are used to classify the tumor because they require percentage of training data to state the parameters "means and variances of the variables" necessary for classification. Extremely fast at classifying unknown records, able to handle both continuous and discrete features, robust to the effect of outliers, provide a clear indication of which fields are most important for prediction, assumption of independence of attributes is too constraining, results/output are incomprehensible, no standardized way for dealing with multi-class problems and fundamentally a binary classifier. The details of the proposed algorithm steps are illustrated.

The proposed algorithm steps:

```
Functions and variables:
X-The instance data,
Y-The list of labels (features)
D[X, Y]-Dataset
K-kernel matrix,
k-number of clusters,
t-number of iterations,
τ-thershold value)
E-Evaluation function (NB, J48, MLP)
Input
D[X, Y]
Output:
 $D_{new}$
Step 1:
        $D_{cls}$ [X, Y] = KKM (D [X, Y], K, k, t)
        $cls_0$ [X, Y] = filter-example $D_{cls}$ [X, Y], p-¯str)
        $cls_1$ [X, Y] =  filter-example $D_{cls}$[X, Y], p-¯str)
Step 2:
        switch (n) :
              case 0:
                  $cls_{new0}$ [X, $Y_{new}$] = CFS ($cls_0$ [X, Y])
                  $cls_{new1}$ [X, $Y_{new}$] = CFS ($cls_0$ [X, Y])
             case 1:
                  $cls_{new0}$ [X, $Y_{new}$] = Relief-F ($Cls_0$ [X, Y],τ)
                  $cls_{new1}$ [X, $Y_{new}$] = Relief-F ($Cls_1$ [X, Y],τ)
             case 2:
                 $cls_{new0}$ IG ($cls_0$ [X, $Y_{new}$], τ)
                $cls_{new1}$ IG ($cls_1$ [X, $Y_{new}$], τ)
 Step 3:
       $D_{new}$ = E ($cls_{new0} \cup cls_{new1}$)
```

The details of the proposed algorithm steps are illustrated in. The proposed method steps are:

Use kernel K-means cluster to divide dataset for two clusters

$$D_{cls}[X, Y] = KKM(D[X, Y], K, k, t)$$

Where:

X = The instance data
Y = The list of labels (features),
K = Kernel matrix: number of clusters, number of iterations. Use filter example for taking each cluster separately and save as dataset independent.

cls_0 [X, Y] = filter-example D_{cls}[X, Y], p-¯str)
cls_1 [X, Y] = filter-example D_{cls}[X, Y], p-¯str)

Where: p-str is parameter string to determine which cluster value filter choose either cluster-0 or cluster-1.

Execute one of feature reduction models (IG,CFS, Relief-F) on each cluster.in case filter evaluation (IG, Relief-F) uses search model (Ranker) is need to determine threshold value (τ) but CFS evaluation uses search model (Best First).

Collection two new clusters in dataset and application classifier (NB, MLP, J48) on new dataset.

RESULTS AND DISCUSSION

The experiments are calculated with following system: Intel core i 3, 4 GB Ram, 500 GB hard drive a Windows 7(64 bit) operating system. The proposed method is executed using RapidMiner and Weka and java "NetBeans 8), Weka is a java based tool used in the field of machine learning and data mining. The input to the system is given as attribute relation file format "ARFF" file. Rapid miner is a computer program platform that provides a combined environment for machine, text mining and data mining.

The training information WDBC "Wisconsin diagnostic breast cancer" dataset which contain 32 features described in which are taken from UCI machine learning repository. Dataset consists of 569 instances divided to two classes are 357 benign and 212 malignant (Frank and Asuncion, 2010).

Features of WDBC dataset:

#: Features:
1: ID number
2: Diagnosis (M = malignant, B = benign)
3.32: Ten real-valued features are computed for each cell nucleus
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeterd)
d) areae
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2/area-1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)i) symmetry
j) fractal dimension ("coastline approximation"-1)

The proposed method is evaluated using five parameters: accuracy, sensitivity, specificity, mean square error and time to build model of the classifier. "ROC" curves are measures that describe statistically the classifier discrimination in case of positive and negative class and visualize the classification process in those cases. These measures are based on some principles which could be defined them as follows:

- True Negature (TN); It refers to the negative instances which are successfully classified by the classifier
- True Positive (TP); It refers to the positive instances which are successfully classified by the classifier
- False Positive (FP); It refers to the negative instances which are unsuccessfully classified as positive
- False Negative (FN); It refers to the positive instances which are unsuccessfully classified as negative

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Sensitivity}\left(\text{TP Rate}\right) = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{14}$$

In this study, the empirical results of the developed breast cancer detection framework with clustering and feature selection are depicted and making comparisons

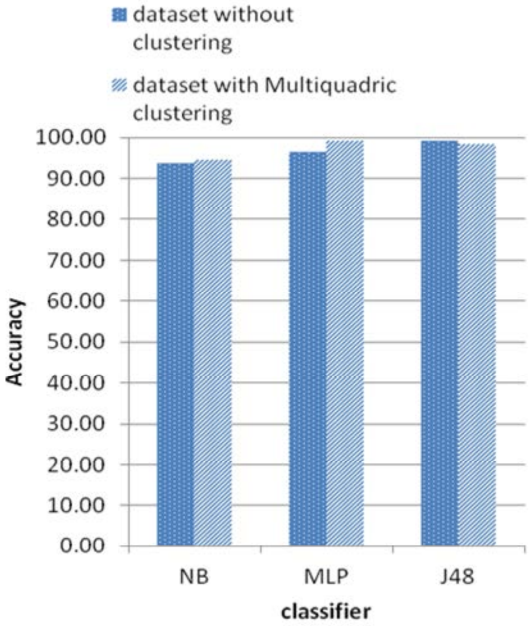

Fig. 2: Accuracy of dataset without clustering and with Multiquadric clustering based on classifiers (NB, MLP, J48) with feature selection IG

Table 1: Cluster instance using multiquadric kernel

		Diagnosis	
Cluster no.	Instances	Benign	Malignant
Cluster_0	488	314	174
Cluster_1	81	43	38

Table 2: Cluster instance using radial kernel

		Diagnosis	
Cluster no.	Instances	Benign	Malignant
Cluster_0	290	195	95
Cluster_1	279	117	162

Table 3: Threshold and search methods

Feature selection	Search methods	Threshold
Relief-F	Ranker	0.06
IG	Ranker	0.40
CFS	Best-irst	-

with other methods existent in the scenario. The proposed method uses k-means algorithm by using rapid miner environment for divided dataset to two clusters twice once using multiquadric kernel as shown in Table 1 number instance for each cluster and the second using radial kernel as shown in Table 2-3 shows threshold values and search methods used with feature selection.

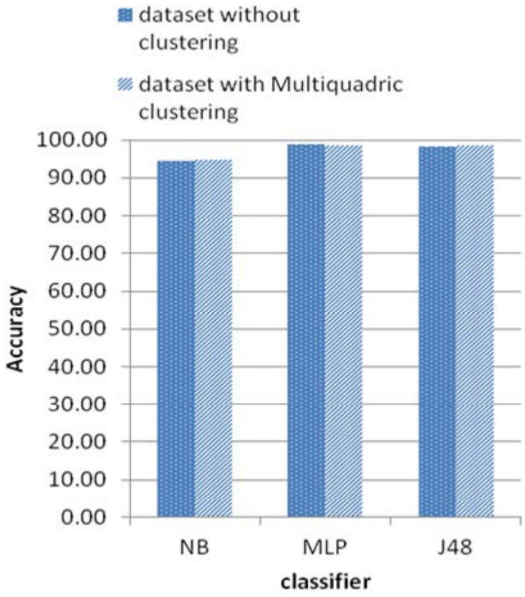

Fig. 3: Accuracy of dataset without clustering and with Multiquadric clustering based on classifiers (NB, MLP, J48) with feature selection relief-F

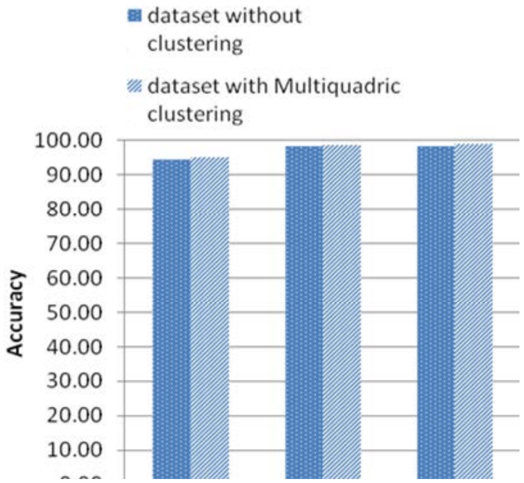

Fig. 4: Accuracy of dataset without clustering and with Multiquadric clustering based on classifiers (NB, MLP, J48) with feature selection CFS

As described in the proposed method, two different ways are used to divide data set into two clusters using kernel k-mean algorithm. By applying the two types of kernels, the first way is by multiquadric kernel and the other by radial kernel. Figure 2 describes the result of accuracy after applied feature selection (IG) on dataset without clustering and dataset with clustering using multiquadric kernel based on classifiers naive bayes, multilayer perceptron and J48. The same comparison describes in Fig. 3 and 4 but applied feature selection relief-F and CFS, respectively.

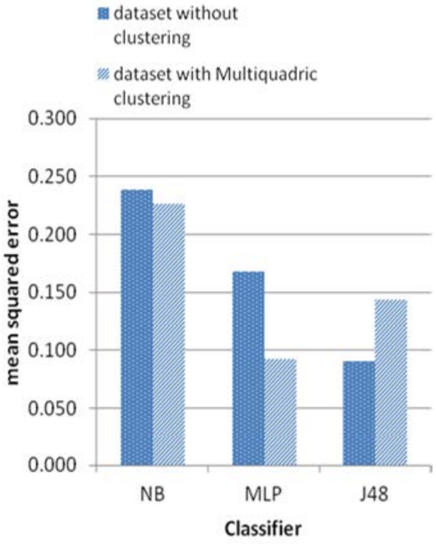

Fig. 5: Mean squared error of dataset without clustering and dataset with Multiquadric clustering based on classifiers (NB, J48) after applied IG

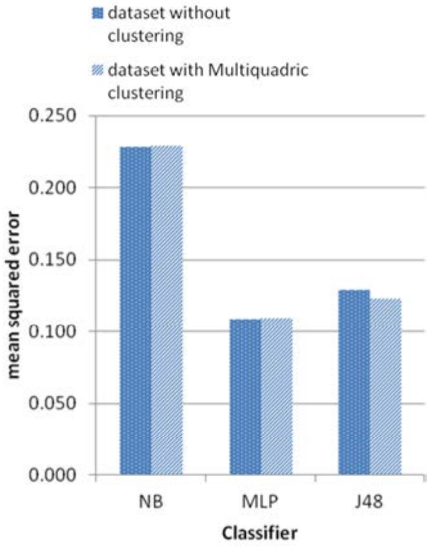

Fig. 6: Mean squared error of dataset without clustering and dataset with Multiquadric clustering based on classifiers (NB, MLP, J48) after applied relief-F

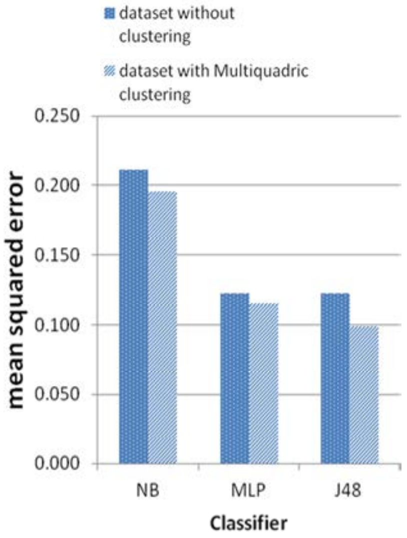

Fig. 7: Mean squared error of dataset without clustering and dataset with Multiquadric clustering based on classifiers (NB, MLP, J48) after applied CFS

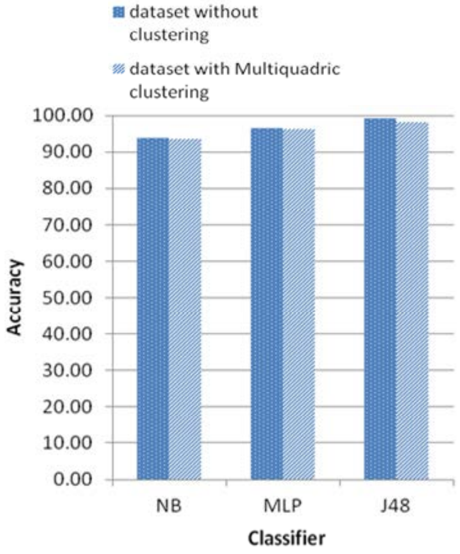

Fig. 8: Accuracy of dataset without clustering and dataset with Radial clustering based on classifiers (NB, MLP, J48) after applied IG

From the previous three figures, it is noticed that the only two cases aren't improved, First case is in Fig. 2 with J48 by using feature selection IG. The second case is in Fig. 3 with MLP by using feature selection Relief-F. The accuracy of the other cases is improved by using the proposed method.

Figure 5 and 6 describes the result of error after applied feature selection (IG) on dataset without clustering and dataset with clustering using multiquadric kernel based on classifiers Naive bayes, MLP and J48. The same comparison describes in Fig. 7 and 8 but applied feature selection relief-F and CFS, respectively.

From the previous three figures, it is noticed that the only one case in Fig. 5 isn't decreased with J48 by using feature selection IG. The mean square error of the other cases is decreased by using the proposed method.

Table 4: Sensitivity specificity and time measurements using multiquadric kernel

		Sensitivity Avg (TPRate)		Specificity Avg (FPRate)		Time to build model (sec)	
Classifier	Feature selection	Dataset without clustering	Dataset with clustering	Dataset without clustering	Dataset with clustering	Dataset without clustering	Dataset with clustering
NB	Relief-F	0.944	0.946	0.068	0.063	0.02	0.02
	IG	0.937	0.944	0.078	0.064	0.03	0.02
	CFS	0.944	0.953	0.072	0.059	0.03	0.03
MLP	Relief-F	0.989	0.986	0.018	0.014	0.83	1.42
	IG	0.965	0.993	0.034	0.012	0.83	1.22
	CFS	0.982	0.986	0.022	0.018	0.95	0.84
J48	Relief-F	0.982	0.984	0.026	0.027	0.08	0.05
	IG	0.991	0.977	0.015	0.038	0.06	0.09
	CFS	0.984	0.989	0.025	0.016	0.11	0.08

Table 5: Sensitivity, specificity and time measurements using radial kernel

		Sensitivity Avg (TPRate)		Specificity Avg (FPRate)		Time to build model (sec)	
Classifier	Feature selection	Dataset without clustering	Dataset with clustering	Dataset without clustering	Dataset with clustering	Dataset without clustering	Dataset with clustering
NB	Relief-F	0.944	0.946	0.068	0.065	0.02	0.03
	IG	0.937	0.935	0.078	0.087	0.03	0.03
	CFS	0.944	0.946	0.072	0.071	0.03	0.02
MLP	Relief-F	0.989	0.991	0.018	0.015	0.83	0.86
	IG	0.965	0.961	0.034	0.036	0.83	0.88
	CFS	0.982	0.986	0.022	0.018	0.95	0.84
J48	Relief-F	0.982	0.984	0.026	0.027	0.08	0.06
	IG	0.991	0.982	0.015	0.030	0.06	0.05
	CFS	0.984	0.986	0.025	0.022	0.11	0.06

Table 6: Some metrics of two clusters using kernel k-means(multiquadric) with three classifiers

		Accuracy		Mean squared error		Sensitivity Avg (TPRate)		Specificity Avg (FPRate)	
Classifier	Feature selection	Cluster 0	Cluster 1	Cluster 0	Cluster 1	Cluster 0	Cluster 1	Cluster 0	Cluster 1
NB	Relief-F	94.26	92.590	0.22	0.24	0.943	0.926	0.075	0.075
	IG	94.06	92.590	0.24	0.25	0.941	0.926	0.074	0.075
	CFS	96.11	98.770	0.18	0.13	0.961	0.988	0.050	0.014
MLP	Relief-F	98.97	100.00	0.10	0.00	0.990	1.000	0.018	0.000
	IG	97.13	100.00	0.15	0.00	0.971	1.000	0.049	0.000
	CFS	98.36	100.00	0.12	0.00	0.984	1.000	0.027	0.000
J48	Relief-F	98.77	100.00	0.11	0.00	0.988	1.000	0.020	0.000
	IG	98.77	100.00	0.11	0.00	0.988	1.000	0.020	0.000
	CFS	98.77	100.00	0.11	0.00	0.988	1.000	0.020	0.000

Table 4 displays the results of sensitivity and specificity before and after applied clustering using k-means (multiquadric kernel). It is calculated on two measurements for three classifiers naïve bayes, MLP and J48.

From the previous Table 1, it is noticed that the only two cases measuring sensitivity aren't improved by using multiquadric kernel clustering. The first case is MLP using feature selection relief-f and the second case is J48 using feature selection IG. Furthermore, there are two cases measuring specificity that aren't decreased. The two cases use J48 by using feature selection relief-f and IG. The time to build the model is measured to achieve the aim of the proposed method of reducing training time. The total improvement of using the proposed method in all cases is 77.8%.

Figure 8 describes the result of accuracy after applied feature selection (IG) on dataset without clustering and dataset with clustering using radial kernel based on classifiers naïve bayes, MLP and J48. The same comparison describes in Fig 9 and 10 but applied feature selection relief-F and CFS, respectively Fig. 9 and 10.

From the previous three figures, it is noticed that the three cases aren't improved, all in Fig 8 by using feature selection IG. The accuracy of the other cases is improved by using the proposed method.

Figure 11 describes the result of error after applied feature selection (IG) on dataset without clustering and dataset with clustering using radial kernel based on classifiers naïve bayes, Multilayer perceptron and J48. The same comparison describes in Fig. 12 and 13 but applied feature selection relief-F and CFS respectively. From the previous three figures, it is noticed that the five cases aren't decreased. The mean square error of the other cases is decreased by using the proposed method. Table 5 displays the results of sensitivity, specificity and time before and after applied clustering using k-means (Radial kernel) that is calculated for three classifiers naïve bayes, multilayer perceptron and J48. From the previous Table 6, it is noticed that the three cases aren't improved

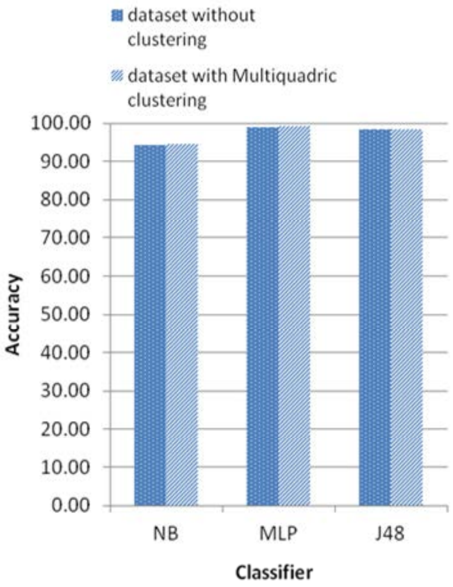

Fig. 9: Accuracy of dataset without clustering and dataset with Radial clustering based on classifiers (NB, MLP, J48) after applied relief-F

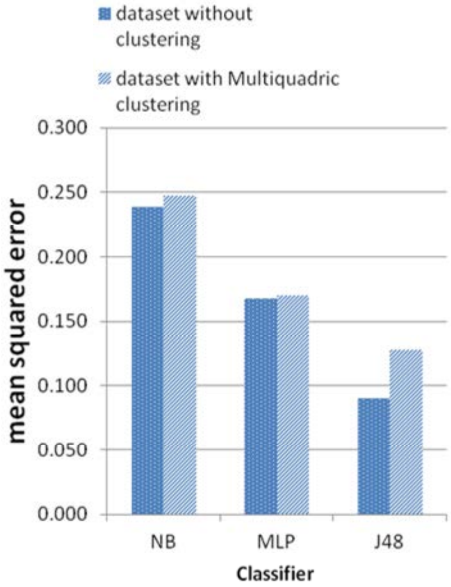

Fig. 11: Mean squared error of dataset without clustering and dataset with Radial clustering based on classifiers (NB, MLP, J48) after applied IG

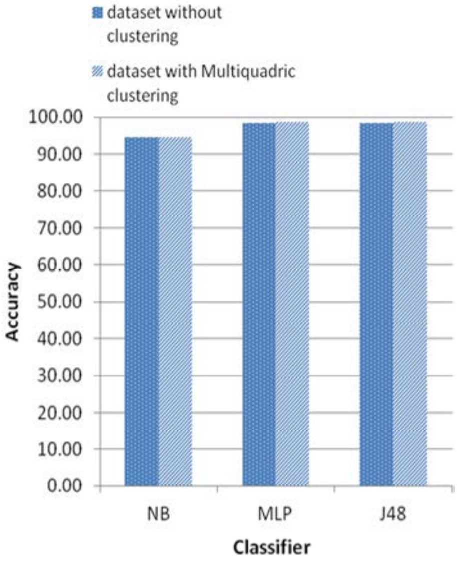

Fig. 10: Accuracy of dataset without clustering and dataset with Radial clustering based on classifiers (NB, MLP, J48) after applied CFS

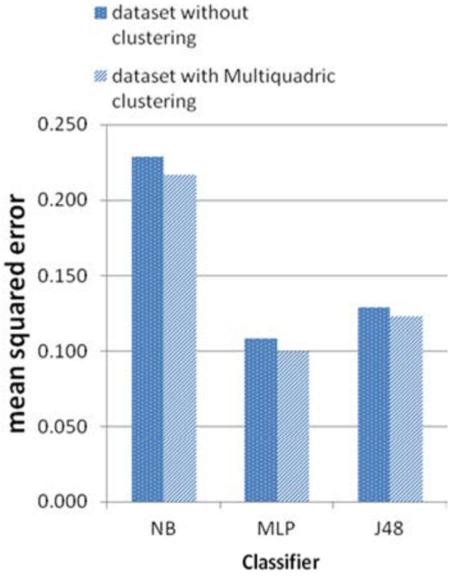

Fig. 12: Mean squared error of dataset without clustering and dataset with Radial clustering based on classifiers (NB, MLP, J48) after applied relief-F

proportion of true positives with three classifiers by using feature selection IG. The sensitivity of the other cases is improved by using the proposed method.

As well it is noticed that the four cases aren't decreased with three classifiers by using feature selection IG and also with J48 by using Relief-F. The specificity of the other cases is decreased by using the proposed method. The time to build model of the most cases is reduced by using the proposed method. Table 6 describes the results after applied three feature selection

Table 7: Some metrics of two clusters using kernel k-means (radial) with three classifiers

		Accuracy		Mean squared error		Sensitivity Avg (TPRate)		Specificity Avg (FPRate)	
Classifier	Feature selection	Cluster 0	Cluster 1	Cluster 0	Cluster 1	Cluster 0	Cluster 1	Cluster 0	Cluster 1
NB	Relief-F	94.48	93.90	0.22	0.21	0.945	0.939	0.081	0.061
	IG	93.45	93.91	0.24	0.22	0.934	0.939	0.102	0.061
	CFS	95.52	95.34	0.20	0.20	0.955	0.953	0.07	0.050
MLP	Relief-F	98.97	98.92	0.11	0.10	0.990	0.989	0.021	0.010
	IG	96.21	95.70	0.16	0.17	0.962	0.957	0.035	0.038
	CFS	98.97	98.92	0.11	0.11	0.990	0.989	0.021	0.013
J48	Relief-F	98.62	97.85	0.12	0.14	0.986	0.978	0.028	0.023
	IG	98.62	96.77	0.11	0.18	0.986	0.968	0.028	0.038
	CFS	98.28	97.85	0.13	0.14	0.983	0.978	0.03	0.023

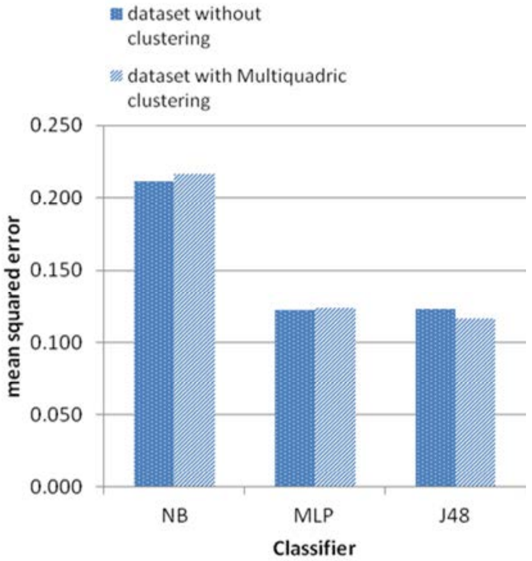

Fig. 13: Mean squared error of dataset without clustering and dataset with Radial clustering based on classifier (NB, MLP, J48) after applied CFS

(IG, relief-F, CFS) respectively on two clusters (divided by using kernel k-means (multiquadric kernel)) before combined them together as subset and measured different measurements for three classifiers Naive bayes, Multilayer perceptron and J48 by weka. Table 7 describes the results after applied three feature selection (IG, relief-F, CFS), respectively on two clusters (divided by using kernel k-means (radial kernel)) before combined them together as subset and measured different measurements for three classifiers Naive bayes, Multilayer perceptron and J48 by weka.

CONCLUSION

In this study we proposed a technique based on machine learning to detect the breast tumor. The hybrid method based on using K- means clustering with feature selection. The clustering has used radial and multiquadric. The feature selection has used relied-f, CFS and IG. It also used the classification to evaluate the system by applying NB, J48 and MLP. There are two experiments. The first one has evaluated the proposed method twice once without clustering and the other with clustering (Multiquadric and Radial). The second one has evaluated the two clusters regularly. The accuracy is improved in most cases with using clustering. The error is also reduced in most cases. The sensitivity, specificity and time to build model has introduced satisfied results. The overall accuracy improvement is almost 77.8% by using Multiquadric clustering. The overall accuracy improvement is almost 66.7% by using radial clustering.

REFERENCES

Aloraini, A., 2012. Different machine learning algorithms for breast cancer diagnosis. Int. J. Artif. Intell. Appl., 3: 21-30.

Badriyah, T., J.S. Briggs and D.R. Prytherch, 2012. Decision trees for predicting risk of mortality using routinely collected data. Int. J. Soc. Hum. Sci., 4: 1-4.

Chan, K., T.W. Lee, P.A. Sample, M.H. Goldbaum and R.N. Weinreb *et al.*, 2002. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. IEEE. Trans. Biomed. Eng., 49: 963-974.

Cristianini, N. and J.S. Taylor, 2000. An Introduction to Support Vector Machines and Other Kernel based Learning Methods. Cambridge University Press, Cambridge, UK., ISBN-13: 9780521780193, Pages: 189.

Ferreira, M.R. and D.F.D.A. Carvalho, 2014. A kernel k-means clustering algorithm based on an adaptive Mahalanobis kernel. Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), July 6-11, 2014, IEEE, Joao Pessoa, Brazil, ISBN:978-1-4799-1484-5, pp: 1885-1892.

Frank, A. and A. Asuncion, 2010. UCI machine learning repository. http://archive.ics.uci.edu/ml/.

Hall, M.A. and L.A. Smith, 1997. Feature subset selection: A correlation based filter approach. Master Thesis, University of Waikato, Hamilton, New Zealand.

Hamad, D. and P. Biela, 2008. Introduction to spectral clustering, information and communication technologies: From theory to applications. Proceedings of the ICTTA 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, April 7-11, 2008, IEEE, France, ISBN:978-1-4244-1751-3, pp: 1-6.

Han, W., J. Dong, Y. Guo, M. Zhang and J. Wang, 2011. Identification of masses in digital mammogram using an optimal set of features. Proceedings of the 2011 IEEE 10th International Conference on Trust Security and Privacy in Computing and Communications, November 16-18, 2011, IEEE, Changchun, China, ISBN:978-1-4577-2135-9, pp: 1763-1768.

Jamuna, K.S., S. Karpagavalli, M.S. Vijaya, P. Revathi and S. Gokilavani *et al*., 2010. Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques. Proceedings of the 2010 International Conference on Advances in Computer Engineering, June 20-21, 2010, IEEE, Coimbatore, India, ISBN:978-1-4244-7154-6, pp: 312-315.

Kira, K. and L.A. Rendell, 1992. The feature selection problem: Traditional methods and a new algorithm. Proceedings of the Tenth National Conference on Artificial Intelligence, July 12-16, MIT Press, San Jose, CA. Cambridge, MA., pp: 129-134.

Kononenko, I., E. Simec and R.M. Sikonja, 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell., 7: 39-55.

Lavanya, D. and D.K.U. Rani, 2011. Analysis of feature selection with classification: Breast cancer datasets. Indian J. Comput. Sci. Eng., 2: 756-763.

Liu, H. and H. Motoda, 2008. Computational Methods of Feature Selection. Chapman & Hall, New York, USA.,.

Liu, R., X. Sun, L. Jiao and Y. Li, 2012. A comparative study of different cluster validity indexes. Trans. Inst. Measur. Control, 34: 876-890.

Manikandan, R., P. Swaminathan and R. Sujitha, 2013. A study on Vlsi physical design specific issues. J. Appl. Sci. Res., 9: 100-103.

Montazeri, M., M. Montazeri, M. Montazeri and A. Beigzadeh, 2016. Machine learning models in breast cancer survival prediction. Technol. Health Care, 24: 31-42.

Mulyono, N.B. and Y. Ishida, 2014. Clustering inventory locations to improve the performance of disaster relief operations. Procedia Comput. Sci., 35: 1388-1397.

Navada, A., A.N. Ansari, S. Patil and B.A. Sonkamble, 2011. Overview of use of decision tree algorithms in machine learning. Proceedings of the IEEE Conference on Control and System Graduate Research Colloquium (ICSGRC), June 27-28, 2011, IEEE, Pune, India, ISBN:978-1-4577-0337-9, pp: 37-42.

Novakovic, J., 2009. Using information gain attribute evaluation to classify sonar targets. Proceedings of the 17th Conference on Telecommunications forum TELFOR, November 24-26, 2009, Megatrend University, Belgrade, Serbia, pp: 24-26.

Osareh, A. and B. Shadgar, 2010. Machine learning techniques to diagnose breast cancer. Proceedings of the 2010 5th International Symposium on Health Informatics and Bioinformatics (HIBIT), April 20-22, 2010, IEEE, Ahvaz, Iran, ISBN:978-1-4244-5968-1, pp: 114-120.

Patil, T.R. and S.S. Sherekar, 2013. Performance analysis of naive bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl., 6: 256-261.

Press, W.H., 1988. Numerical Recipes in C. Cambridge University Press, Cambridge, UK.,.

Thein, H.T.T. and K.M.M. Tun, 2015. An approach for breast cancer diagnosis classification using neural network. Adv. Comput., 6: 1-11.

Tzortzis, G. and A. Likas, 2008. The global kernel k-means clustering algorithm. Proceedings of the 2008 IEEE International Joint Conference on Neural Networks and World Congress on Computational Intelligence, June 1-8, 2008, IEEE, Greece, ISBN:978-1-4244-1820-6, pp: 1977-1984.

Vilan, J.V., J.A. Fernandez, P.G. Nieto, F.S. Lasheras and D.C.F.J. Juez *et al*., 2013. Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir Northern Spain. Water Resour. Manage., 27: 3457-3476.

Win, S.L., Z.Z.F. Htike and I.A. Noorbatcha, 2014. Cancer recurrence prediction using machine learning. Int. J. Comput. Sci. Inf. Technol., 6: 11-20.

Ying, K., A. Ameri, A. Trivedi, D. Ravindra and D. Patel *et al*., 2015. Decision tree-based machine learning algorithm for in-node vehicle classification. Proceedings of the Conference on Green Energy and Systems Conference, November 9-9, 2015, IEEE, Long Beach, California, ISBN:978-1-4673-7263-3, pp: 71-76.

Zheng, B., S.W. Yoon and S.S. Lam, 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst. Appl., 41: 1476-1482.