Research Journal of Applied Sciences, Engineering and Technology 10(1): 49-55, 2015
DOI:10.19026/rjaset.10.2553
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 30, 2014 Accepted: January 11, 2015 Published: May 10, 2015

Research Article

A Novel Hybrid System for Diagnosing Breast Cancer Using Fuzzy Rough Set and LS-SVM

[1]R. Jaya Suji and [2]S.P. Rajagopalan
[1]Sathyabama University,
[2]GKM College of Engineering and Technology, Chennai, India

Abstract: With fast development of medical diagnosis technologies, the filtering of the entire relevant feature and time consuming task are challenging tasks. For effective feature selection and reducing the time consuming, we propose a new hybrid system for diagnosing the breast cancer. The proposed hybrid system is the combination of CFRSFS, K-Means Clustering and Least Square Support Vector Machine (LS-SVM). In this hybrid system, we propose a new feature selection algorithm called Correlation based Fuzzy Rough Set Feature Selection (CFRSFS) algorithm for effective initial feature selection process. Moreover, K-Means clustering algorithm has been used for enhancing the feature selection process based on the factors of the selected features in similar manner of the existing hybrid system. Finally, LS-SVM algorithm is also used for classifying the feature selected breast cancer dataset. The experiments have been conducted for evaluating the proposed system using WDBC Data set. The obtained results show that the performance of the proposed system classification accuracy is 99.54%.

Keywords: Cancer diagnosis, data mining, k-means clustering, least square support vector machine

INTRODUCTION

Cancer disease has become a severe health problem in the world, as in recent times the number of deaths also increasing rapidly. In United States, the number of cancer deaths reached at 577,190 (Siegel *et al*., 2012). The uncontrolled growth of cells in the body is known as cancer. From that, the breast cancer is the erratic growth of cells that originate in the breast tissue. Tumor is a group of extra tissues and these tumors can be either cancerous or non-cancerous. Cancerous tumors affect healthy tissues initially in the body and this brings it to death.

Medical science is facing a very serious issue of tumors diagnosis. With the advent and rapid growth of information technology, the researchers introducing new applications and tools to the society in every day means to get group of meaningful tumor data and core medical information on cancer research. Even though, to learn about cancer features is challenging task today from increasing the huge volume of cancer cases. Therefore, data analysis based prediction system is a useful tool for medical doctors when making decisions on cancer diagnosis. Mining techniques have been used to analyze all kinds of cancer diagnoses include breast cancer. The mining techniques are used to create new diagnosis systems to detect the serious disease like cancer early or start the treatment for respective disease. The dimensionality is a big problem when we have huge volume of data for analyzing due to the presence of redundant and noisy features and these affects the performances of the medical systems (Zheng *et al*., 2014).

Data preprocessing is playing a major role in any system like medical diagnosis systems which are developed based on the classification methods. The data can be preprocessed either the way of feature reduction or selection. Feature selection selects the necessary features and also eliminates the unnecessary features from the dataset. This process is used for improving the performance of the classification algorithms. Generally, this technique can be classified into two methods namely filter methods and wrapper methods. From these, the wrapper methods are used to optimize a predictor as part of the selection process. Filter methods depends on the general characteristics of the given input training data to select optimal features for any predictor independently (Ganapathy *et al*., 2013).

Clustering methods are used to investigative data analysis technique and it tries to make it many small groups of given data set into dissimilar groups. The classes of every group are more similar to one another than those belonging to various groups. Generally, the clustering techniques are categorized into two namely supervised methods and unsupervised methods. The supervised clustering method activates by human and unsupervised clustering method used to detect the underlying structure in the data set for classification.

Corresponding Author: R. Jaya Suji, Sathyabama University, Chennai, India

The unsupervised clustering techniques are very famous because of they do not need much knowledge about the data sets (Ganapathy *et al*., 2012a, b).

Classification is used to train a model called classifier from a set of records called training and then to categorize a test records into one of the classes using the trained model known as testing (Ganapathy *et al*., 2013). Classifier systems in the recent times have shown its importance in the medical diagnosis. The evaluation of the data obtained from the patients and the decision of the expert are very important factors in diagnosis of any disease by the expert. However, to assist the expert to make decisions, expert systems and various artificial intelligence techniques help to the great extent for classification. Classification systems will be helpful to minimize the possible errors and also provides the detailed analysis on medical data efficiently (Zheng *et al*., 2014).

In this study, a new hybrid system for diagnosing the breast cancer is proposed. This hybrid system is the combination of CFRFS algorithm, K-Means Clustering (Zheng *et al*., 2014) and LS-SVM (Polat and Günes, 2007). Here, a new CFRFS algorithm is proposed for initial feature selection process. Moreover, the K-Means Clustering algorithm is used for enhancing the feature selection process further. Finally, we used LS-SVM algorithm for classifying the feature selected input data. This proposed hybrid medical diagnosis system helps to improve the diagnosis process effectively with less time.

LITERATURE REVIEW

Data mining techniques are used to extract the particular patterns from huge volume of data which includes noisy data (Fayyad *et al*., 1996). It uses the various fields such as statistical analysis, machine learning techniques and artificial intelligence for analyzing the data (Venkatadri and Lokanatha, 2011). The major tasks of this data mining are clustering and classification. The main task of the classification is to find the common features which are useful to solve a problem and classified as many classes (Chen *et al*., 1996). It is related to clustering problems. In classification problems, the label of each class is a discrete and known category, while the label is an unknown category in clustering problems (Xu and Wunsch, 2005). Clustering problems were considered as unsupervised classification problems (Dubois and Prade, 1990). The clustering process is summarizes the data for the particular patterns without existing class labels. Generally, the breast cancer is treated as a classification problem in which used to search for the particular classification problem to classify cancerous and noncancerous tumors.

Support Vector Machine (SVM) is an effective statistical learning model for classification (Jain *et al*., 1999; Cortes and Vapnik, 1995) and it support vectors used to set the margin as decision boundaries between two different classes. It is fully functioning based on a linear machine of a high dimensional feature space, nonlinearly which is related to the input space and it has allowed the development of the fast training methods, even with a huge number of input features and training data sets (Cortes and Vapnik,1995). The Least Square Support Vector Machine (LS-SVM) is an enhanced version of SVM and it has been introduced by (Polat and Günes, 2007) for breast cancer diagnosis system. The authors used the WDBC dataset and achieved 98.53% accuracy from their experiments. It is used a set of linear equations for training the dataset for solving the quadratic optimization problem on huge volume of data.

Ferreira and Figueiredo (2012) introduced a new unsupervised method for feature discretization which is based on the Linde-Buzo-Gray algorithm. Their algorithm has two methods for allocating a feature number of bits and second uses a fixed number of bits per feature. Both methods try to adequate representations for supervised and unsupervised feature selection methods with different kinds of classifiers. Moreover, they also introduced an efficient feature selection method for redundancy analysis and to choose an adequate number of features for the particular problem. Li *et al*. (2014) proposed a rough set theory based technique named Total Mean Distribution Precision (TMDP) for selecting the partitioning variable. Based on this technique they derived a new clustering algorithm called Maximum Total Mean Distribution Precision (MTMDP) for categorize the input dataset. This is a powerful clustering algorithm for handling uncertainty during the process of data categorization. Authors proved from their experiments this method also can be implemented to analyze the particular grouped categorical data for producing better clustering results.

Ganapathy *et al*. (2012) introduced a new immune genetic algorithm based clustering technique for classification. Authors achieved better classification accuracy than the existing classifiers when handling the online dataset like social networks. Chen (2014) proposed a hybrid model is based on that uses the combination of cluster analysis and feature selection for analyzing breast cancer diagnoses. This model provides a better way of selecting the subset of salient features for performing clustering and also uses the most existing features depended methods for clustering. In particular, their methods select the salient features to identify exact cluster techniques using quantitative measurements and comparisons.

Zheng *et al*. (2014) presented a hybrid system called the K-SVM based on the recognized feature patterns which is the combination of K-means clustering algorithm and SVM for cancer diagnosis.

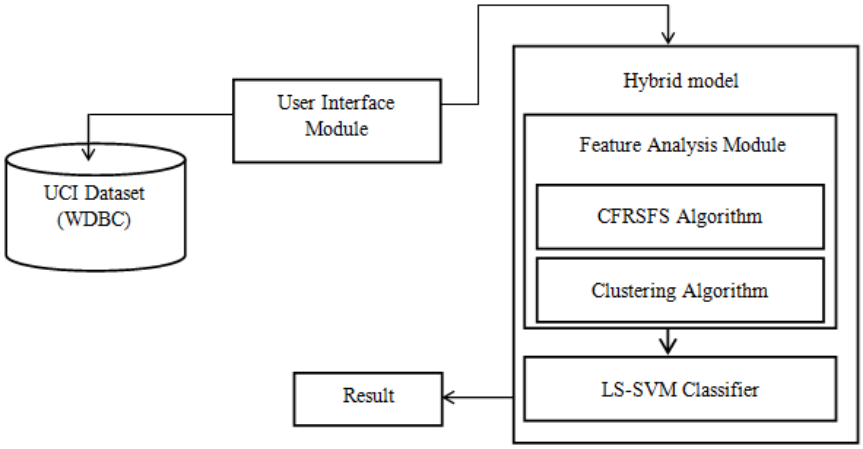

Fig. 1: System architecture

Here, the K-means clustering algorithm is used for recognizing the cancerous patterns and obtaining the noncancerous patterns of the breast cancer dataset. These obtained patterns are reconstructed as the new tumor features for training phase for the solution of the particular problem. The feature extraction and feature selection provides effective feature set for the machine learning algorithms to train the classifier.

In this study, we proposed a new hybrid system is the combination of a newly proposed feature selection algorithm called Correlation based Fuzzy Rough Set Feature Selection (CFRFS) algorithm, K-Means Clustering algorithm (Zheng *et al*., 2014) and LS-SVM (Polat and Günes, 2007) for diagnosing the breast cancer.

System architecture: The architecture of the system proposed in this study consists of four major modules namely, user interface module, feature analysis module, LS-SVM classifier module and result module as shown in Fig. 1.

The User interface module collects the WDBC data from the UCI repository machine learning dataset. This data is sent to the hybrid module for preprocessing and classification. The feature analysis module analyzes the features by using the proposed CFRSFS algorithm and K-Means clustering algorithm. This module selects only the valuable attributes from the data set using projection. Moreover, the basic data preprocessing steps are carried out for performing effective preprocessing. This module classified the data initially based on the properties of dataset. After that, these selected features of data can be forwarded into the classification module for the further classification. Finally, results can be produced which can be accessed by the user interface.

METHODOLOGY

In this study, a new hybrid system is proposed for diagnosing the breast cancer effectively. The proposed hybrid system is the combination of newly proposed feature selection algorithm called Correlation based Fuzzy Rough Set Feature Selection (CFRSFS) algorithm and K-Means Clustering (Zheng *et al*., 2014) and LS-SVM (Polat and Günes, 2007) for diagnosing the breast cancer.

Fuzzy rough set: The following preliminary information about fuzzy rough sets can be found in (Jaisankar *et al*., 2012). Fuzzy set theory introduced by Zadeh (1965) for representing the data uncertainty. Primary concern of fuzzy set is reasoning using natural languages in which words can have ambiguous meanings. First fuzzy rough set idea is proposed by Dubois and Prade (1990) and they defined two approximation functions namely lower and upper approximation as follows.

Let U be the nonempty universal set and R be a Fuzzy binary relation on U, F (U) be the fuzzy power set of U. A fuzzy rough set is a pair (O* (F) , O* (F)) of fuzzy rough set on U such that for every X Є U the upper and lower approximations are defined as follows:

$$\mu_{\emptyset} * (F)F_i = \inf_{x \in U} \min\{F_i(x), F(x)\} \quad (1)$$

$$\mu_{\emptyset*(F)}(F) = \sup_{x \in U} \max\{1 - F_i(x), F(x)\} \quad (2)$$

where, $\emptyset = \{F_1, F_2, \ldots \ldots F_k\}$ is a fuzzy partition derived from R. That $\emptyset$ is the collection of all fuzzy equivalence classes of R. sup and inf indicate the least upper bound (Supermom) and the greatest lower bound (infimum).

The basic properties of Dubois and Prade (1990) rough sets have been extended by them to fuzzy rough sets as follows:

O* (F∪G) ⊇O* (F) ∪O* (G)

O* (F∪G) = O* (F) ∪O* (G)

O* (F∩ G) ⊆O* (F) ∩ O* (G)

O* (F∩ G) = O* (F) ∩ O* (G)

$O^*(Fc) = (O^*(F))^c$

These properties have been used to form fuzzy rules which are used for decision making.

Correlation feature selection: Hall (1998) described a new feature selection method called Correlation based Feature Selection (CFS) for improving the performance of diagnosis system. The worth of the individual features in order to predict the class and the level of inter-correlation among classes is calculated by this method by implementing heuristics. The weightage of a subset of attributes is computed by this method taking in consideration the predictive ability of each feature individually with the degree of redundancy between them.

Correlation based fuzzy rough set feature selection: Correlation based Fuzzy Rough Set Feature Selection (CFRSFS) is a two-level feature reduction algorithm which is the combination of two feature selection algorithms namely Correlation Feature Selection (CFS) Algorithm (Hall, 1998) and Fuzzy Rough Attribute Reduction (FRAR) algorithm (Jensen and Shen, 2004). The purpose of this algorithm is to frame an effective feature subset with useful features for improving the performance of the breast cancer diagnosis system.

Correlation based fuzzy rough set feature selection algorithm:
Input: The set S of all features
Output: F, the set of optimal features
Algorithm:
//Let A be the set of features

Step 1: Initialize F1, F2, F3 to all null set. i.e., F1 = { }, F2 = {}, F3 = {}
Step 2: S is discretized first with the help of equal frequency in five intervals
Step 3: Calculate the Correlation Value (CV) for all the attributes of S using CFS
Step 4: Find the Mean Value (MV) for correlation values
Step 5: For each attribute A_j do Begin
Step 6: for j = 1 to n do

Begin
If CV [A_j] >MV then
Call FRAR ()
F1 = F1 U (A_j)
Else if CV [A_j] = MV then
Call FRAR ()
F2 = F2 U (A_j)
Else
Call FRAR ()
F3 = F3 U (A_j)
End
End

Step 7: List the selected features. i.e., F = {F1 U F2 U F3}
Step 8: Call CFS (F)
Step 9: Output the set that contains the selected features of F

The input data sets are discretized for the purpose of fuzzy rough set by using the frequency with number of intervals 5 in fuzzy logic manner initially before start the feature selection process. Calculate the correlation value for all the attributes with the help of correlation method (Polat and Günes, 2007) then find the mean value for the values. Three subsets have been formed based on the correlation values which are less, greater and equal to the mean value of the correlation then combined all the subsets. CFS has been applied for further attribute reduction for the selected attributes. Finally, produce the list of selected features.

K-means clustering: In this study, we used a K-Means clustering algorithm for enhancing the feature selection process according to the existing hybrid system (Hall, 1998). The K-means algorithm is playing major role in medical diagnosis system and for avoiding empty clusters with less members K constrains are added for reducing the cluster numbers. This K-means algorithm does not require a search method on feature selection instead of providing a good number of samples reductions without applying any feature selection and extraction. Here, a hybrid of Correlation based Fuzzy Rough Feature Selection algorithm, K-means algorithm (Zheng *et al*., 2014) and LS-SVM (Polat and Günes, 2007) is to reduce the features and due to reduce the time taken by LS-SVM for training and testing.

Least square support vector machine: Least Squares-Support Vector Machine (LS-SVM) classifier (Polat and Günes, 2007) is one particular sample of SVM. It could be find the solution for solving a set of linear equations instead of a convex quadratic programming problem on SVMs. The main goal of this algorithm is to find an optimal hyper plane which is separate into various classes. It uses the maximum Euclidean distance of hyper-plane for finding the nearest point. This classifier can be achieved better classification accuracy than the SVM due to find the closest point.

RESULTS AND DISCUSSION

The proposed hybrid system has been implemented using MATLAB (Version 7.12). It is used to avoid the irrelevant and redundant features.

WDBC data set: In this research work, a data set of Wisconsin Diagnostic Breast Cancer (WDBC) (Wolberg *et al*., 1995) from the University of California-Irvine repository has been used for implementation which is used in worldwide by the researchers for breast cancer diagnosis system. It

Table 1: Performance of proposed feature selection algorithm

Data mining algorithms	Full features (accuracy %)	CFRSFS selected features (accuracy %)
Naïve Bayes	90.3509	94.7368
RBF	92.9825	99.1228
SMO	97.3684	97.3684
J48	92.9825	96.4912
Simple cart	92.1053	94.7368

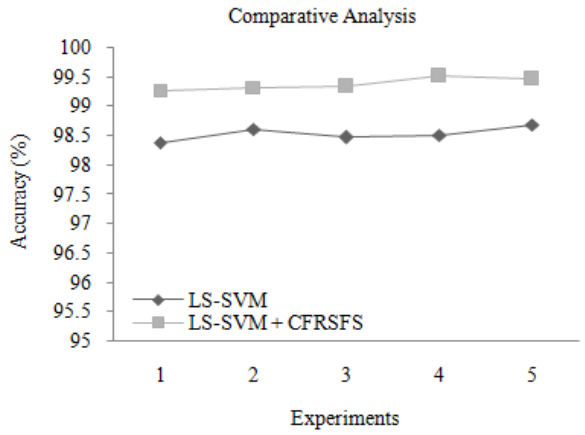

Fig. 2: Comparative analysis between LS-SVM and LS-SVM with CFRSFS

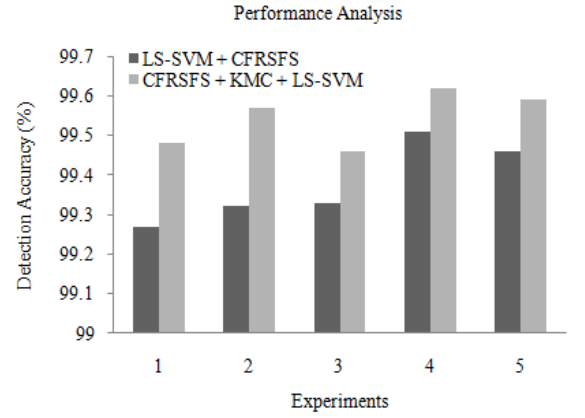

Fig. 3: Performance comparative analyses between LS-SVM with CFRSFS and CFRSFS+KMC+LS-SVM

contains 32 features in 10 varieties for each cell nucleus includes radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The mean value, standard error and maximum value are measured in each category. These measurements are treated as various features in the data set. The removal of different scale effects and normalization of the data is required for training before start the training process. We have collected and used 569 instances for experiments with the diagnosed cancer results.

Experimental results: The proposed CFRSFS algorithm has selected only 8 features dimensions from 32 feature dimensions. The selected features for WDBC dataset is the feature dimension of 11, 17, 22, 24, 25, 26, 28 and 31, respectively. The proposed feature selection algorithm has combined with various data mining algorithms for the different experiment in WEKA tool. Table 1 shows the performance of the proposed feature selection algorithm with the various data mining algorithms.

From Table 1, it can be observed that the performance of the uses of selected features by proposed CFRSFS is better when it is compared with the uses of full features in various standard data mining algorithms.

The K-Means clustering algorithm also performed the feature extraction and selection. Moreover, the performance of the classification and the process of feature extraction have been improved by K-Means clustering algorithm due to eliminating the unnecessary factors of the selected features. Based on the presence of unnecessary factors the selected features also can be reduced. Finally, this combined method has selected only 6 features dimensions from 32 features dimensions in WDBC data. Initially, eight features have been selected by the CFRSFS then after applying the K-Means clustering algorithm reduced the number of features to six.

LS-SVM is provided 98.53% accuracy for breast cancer diagnosis system with WDBC breast cancer data set which is proposed by Polat and Günes (2007) and Ganapathy *et al*. (2013). Figure 2 shows that the comparative analysis between LS-SVM and LS-SVM with CFRSFS.

From Fig. 2, it can be observed that the performance of the proposed feature selection algorithm with LS-SVM is better when it is compared with LS-SVM. The combination of LS-SVM and CFRSFS is provided 99.378% accuracy. It is nearly 1% is higher than which is provided by LS-SVM.

Figure 3 shows that the performance comparative analysis between LS-SVM with CFRSFS and LS-SVM with CFRFS and K-Means Clustering. From Fig. 2, it can be observed that the performance of the proposed method (combination of LS-SVM, CFRSFS and K-Means Clustering) is provided better detection accuracy when it is compared with the combination of LS-SVM and CFRSFS. The proposed method is provided 99.54% accuracy by using less number of features.

Table 2 shows the comparative analysis of various hybrid systems which are proposed by various researchers in the past for diagnosing the breast cancer. The proposed hybrid method provides better accuracy when it is compared to the previous experimental results by Prasad *et al*. (2010). From the computation time perspective, the proposed method reduces the training time significantly by decreasing the number of the input features.

The proposed hybrid method is to recognize the patterns of important features of benign and malignant

Table 2: Results comparison

Method	Number of features used	Detection accuracy (%)
CFRSFS+KMC+LS-SVM	6	99.54
K-SVM (Chen, 2014)	6	97.38
PSO-SVM (Ganapathy *et al*., 2012a)	15	97.37
GA-SVM (Ganapathy *et al*., 2012a)	18	97.19
ACO-SVM (Ganapathy *et al*., 2012a)	17	95.96

tumors. It reconstructed the new input feature with the membership of new patterns based on original data that provides better accuracy than other approaches. Feature selection process filters the unnecessary factors for the diagnosis of cancer. Therefore, the data pre-processing (feature extraction and selection) plays a major role in the cancer diagnoses. The proposed system reduces the number of dimensions of features. This helps to reconstruct the structure of the features for supporting the machine learning algorithm to optimize the classifier. The classifier accuracy is also improved based on the elimination of not important features and unnecessary factors of features.

CONCLUSION AND RECOMMENDATIONS

A new hybrid system has been proposed and implemented in this study for the breast cancer diagnosis. The proposed hybrid system is the combination of CFRSFS, K-Means Clustering and LS-SVM. In this hybrid system, we propose a new feature selection algorithm called Correlation based Fuzzy Rough Set Feature Selection (CFRSFS) algorithm for effective initial feature selection process. CFRSFS selects eight features, these features data only can give as input to the K-Means clustering algorithm for further feature extraction process. We used the K-Means clustering algorithm (Zheng *et al*., 2014) for enhancing the feature selection process based on the factors of the selected features in similar manner of the existing hybrid system. K-means clustering algorithm is used to recognize and obtain the patterns of malignant and benign tumors, respectively. Finally, we used LS-SVM algorithm (Polat and Günes, 2007) for classification. This proposed hybrid system has achieved better detection accuracy (99.54%) when compared with other existing hybrid method. The existing hybrid methods achieved only 97.38, 97.37, 97.19 and 95.96%, respectively of accuracy. Future works in this direction could be the use of effective fuzzy logic for enhancing the power of the decision manager. For reducing the time taken for this hybrid model can use the intelligent agents and the knowledge base system.

REFERENCES

Chen, C.H., 2014. A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Appl. Soft Comput., 20: 4-14.

Chen, M.S., J. Han and P.S. Yu, 1996. Data mining: An overview from a database perspective. IEEE T. Knowl. Data En., 8: 866-883.

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20: 273-297.

Dubois, D. and H. Prade, 1990. Rough fuzzy sets and fuzzy rough sets. Int. J. Gen. Syst., 17: 191-208.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery in databases. Artif. Intell. Mag., 17: 37-54.

Ferreira, A.J. and M.A.T. Figueiredo, 2012. An unsupervised approach to feature discretization and selection. Pattern Recogn., 45: 3048-3060.

Ganapathy, S., P. Yogesh and A. Kannan, 2012b. Intelligent agent based intrusion detection system using enhanced multiclass SVM. Int. J. Comput. Intell. Neurosci., 20(12): 195-202.

Ganapathy, S., K. Kulothungan, P. Yogesh and A. Kannan 2012a. A novel weighted fuzzy c-means clustering based on immune genetic algorithm for intrusion detection. Proc. Eng., 38: 1750-1757.

Ganapathy, S., K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh and A. Kannan, 2013. Intelligent feature selection and classification techniques for intrusion detection in networks: A survey. EURASIP J. Wirel. Comm., 271: 1-16.

Hall, M., 1998. Correlation-based feature selection for machine learning. Ph.D. Thesis, Department of Computer Science, Waikato University, Hamilton, NZ.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surv., 31: 264-323.

Jaisankar, N., S. Ganapathy and A. Kannan, 2012. Intelligent intrusion detection system using fuzzy rough set based C4.5 algorithm. Proceeding of International ACM Conference on Advances in Computing, Communications and Informatics (ICACCI-2012), pp: 596-601.

Jensen, R. and Q. Shen, 2004. Fuzzy-rough attributes reduction with application to web categorization. Fuzzy Set. Syst., 141(3): 469-485.

Li, M., S. Deng, L. Wang, S. Feng and J. Fan, 2014. Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowl-Based Syst., 65: 60-71.

Polat, K. and S. Günes, 2007. Breast cancer diagnosis using least square support vector machine. Digit. Signal Process., 17: 694-701.

Prasad, Y., K. Biswas and C. Jain, 2010. SVM classifier based feature selection using GA, ACO and PSO for siRNA design. Proceeding of the 1st International Conference on Advances in Swarm Intelligence, pp: 307-314.

Siegel, R., D. Naishadham and A. Jemal, 2012. Cancer statistics. Cancer J. Clin., 62: 10-29.

Venkatadri, M. and C.R. Lokanatha, 2011. A review on data mining from past to the future. Int. J. Comput. Appl., 15: 19-22.

Wolberg, W.H., W.N. Street and O.L. Mangasarian, 1995. Image analysis and machine learning applied to breast cancer diagnosis and proganosis. Anal. Quant. Cytol., 17(2): 77-87.

Xu, R. and D. Wunsch, 2005. Survey of clustering algorithms. IEEE T. Neural Networ., 16: 645-678.

Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8: 338-353.

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.