Indian Journal of Science and Technology, *Vol 8(27), DOI: 10.17485/ijst/2015/v8i27/53086, October 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Genetic Algorithm based CFS and Naive Bayes Algorithm to Enhance the Predictive Accuracy

T. Karthikeyan[1] and P. Thangaraju[2*]

[1]Department of Computer Science, PSG College of Arts and Science, Coimbatore – 641014, Tamil Nadu, India
[2] Department of Computer Applications, Bishop Heber College, Tiruchirappalli – 620017, Tamil Nadu, India; thangarajubhc@yahoo.co.in

Abstract

For better classification accuracy in the area of data mining, feature selection techniques are applied to medical data sets. In this work genetic algorithm and Particle Swarm Optimization search techniques and correlation based feature selection is used for evaluation and naive bayes classifier for classification purpose. The hepatitis data set, taken from the UCI machine learning repository, is applied in this work. Accuracy and time is the outcome of the classification model and also various measures like sensitivity, specificity, precision and recall are also calculated.

Keywords: Correlation based Feature Selection, Data Mining, Feature Selection, Genetic Algorithm, Naive Bayes Algorithm, Particle Swarm Optimization

1. Introduction

Feature Selection[1,2] techniques reduces the dimensionality of data or eliminates the irrelevant features and improves the predictive accuracy. The feature selection begins with an empty set of features and generates all possible single feature expansions and the subset with the maximum accuracy is chosen and expanded in the same way by adding single features. The search continues, if the accuracy's subset expansion is maximized, then the search goes to the next best unexpanded subset. Then, the subset with the maximum accuracy will be selected as the reduced feature set[3-8].

This works is based on applying feature selection techniques for hepatitis data set inorder to improve the classification accuracy. This work applies Correlation based Feature Selection (CFS)[9,10] with Genetic Algorithm[9-11] and Particle Swarm optimization[12-14] as searching techniques for the selection of subset and naive bayes algorithm for classification. The comparison study was conducted with other classification algorithms like Sequential Minimal Optimization (SMO), J48, Multi Layer Perceptron (MLP)[15] and Radial Basis Function (RBF).

In[11] proposed a novel feature selection method based on CFS. Initially, the measures of variable to variable and variable to observe were calculated respectively. Then heuristic search method to search the space of variable for selecting informative gene subset was utilized and the subset weight was computed using these measures. Through regression a subset of distinguished genes was obtained. The stratified sampling strategy was presented to obtain the most exposed genes and the classification performance was tested to evaluate the proposed method applies Ten-fold cross-validation for the leukemia, colon cancer and prostate tumor datasets.

In[16] dealt with classification with automatic blood glucose data from patient's glucose meter for the development of decision support systems for gestational diabetes. They used feature selection methods and neural net-

*Author for correspondence

works and decision trees classification algorithm for their research.

In[17] applied CFS using Neighborhood Mutual Information (NMI) and PSO are combined into an ensemble technique. Based on this observation, an efficient gene selection algorithm, denoted by NMICFS-PSO, was developed and several cancer recognition tasks are collected for testing the proposed technique. Further, Support Vector Machine integrated with leave-one-out cross-validation and calculated the classification accuracy.

In[18] developed a feature selection algorithm called Modified Correlation Rough Set Feature Selection (MCRSFS) that predicts both diagnosis and prognosis by compared with several data mining classification algorithms. In their approach features are selected based on rough set with different starting values of reduction in stage one and in stage two features are selected from the reduced set based on the CFS.

2. Correlation based Feature Selection

CFS[9,10,18] is one of the well-known techniques to rank the relevance of features by measuring correlation between features and classes and between features and other features. Given number of features k and classes c, CFS defined relevance of features subset by using Pearson's correlation equation

$$M_S = k\overline{r_{cf}} / \sqrt{k + (k-1)\overline{r_{ff}}} \quad (1)$$

Where M_s is relevance of feature subset, $\overline{r_{cf}}$ is the average linear correlation coefficient between these features and classes and $\overline{r_{ff}}$ is the average linear correlation coefficient between different features.

Normally, CFS adds (forward selection) or deletes (backward selection) one feature at a time. However, in this work we used Genetic search and PSO search algorithms for the good results[19,20].

2.1 Genetic Search Algorithm

Search methods traverse the attribute space to find a good subset and the quality is measured by the attribute subset evaluator through CFS subset evaluator and genetic search is being used as a search method. The parameters of the genetic algorithm are number of generations, population size and the probabilities of mutation and crossover. A member of the initial population generates by specifying a list of attribute indices as a search point. For generating progress reports, every so many generations can be used[20,21]. The simple genetic search strategy is shown below:

Step 1: Start by randomly generating an initial population

Step 2: Calculate e(x) for each member xεP.

Step 3: Define a probability distribution p over the members of P where p(x) ∞ e(x).

Step 4: Choose two population members x and y to produce new population members x' and y'.

Step 5: Apply mutation to x' and y'.

Step 6: Insert x' and y' into P'

Step 7: If |p'| < |P|, go to step 4.

Step 8: Let P ←P'.

Step 9: If there are more generations to process, go to step 2.

Step 10: Return x ε P for which e(x) is highest.

2.2 Particle Swarm Optimization Search

PSO[8] is a stochastic optimization method based on the behavior of swarming animals such as birds and fish. A number of particles, representing potential solutions to the problem are released in the search space of potential solutions. Each particle has a position and a velocity and is free to fly around the search space. The movement is controlled; the particles accelerate towards the position of the best performing particle as well as towards each particle's personal best previous position. The PSO algorithm is governed by a set of rules describing how each particle's position and velocity changes over time[22-26]. The following shows the PSO search strategy.

Step 1: Initialize random positions and velocities for particles, x_i, v_i, i = 1…n

Step 2: Evaluate each particle in the swarm, $x_i \rightarrow f(x_i)$

Step 3: For each particle update the best position and the global best position.

if $f(x_i) > f(p_i) \rightarrow p_i = x_i$

if $f(x_i) > f(p_g) \rightarrow p_g = x_i$

Step 4: For each particle update velocity and position, according to equation (2) and (3)

$$x_{id}^{t+i} = x_{id}^{t} + v_{id}^{t+1} \quad (2)$$

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_{1i} * (p_{id} - x_{id}^{t}) + c_2 * r_{2i} * (p_{gd} - x_{id}^{t}) \quad (3)$$

Step 5: Unless the termination criterion has been met, return to step 2.

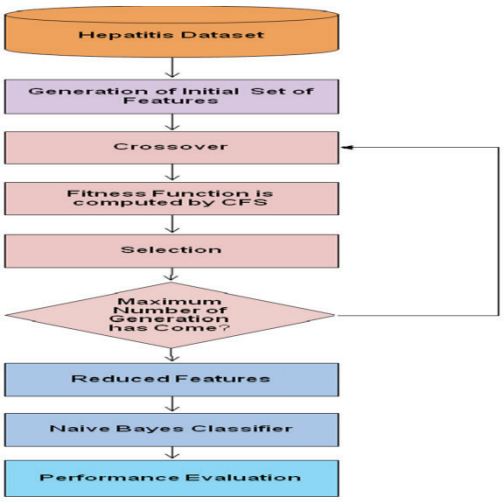

Figure 1. Proposed Methodology for GA based CFS and Naive Bayes Classification.

Table 1. Data set descriptions of hepatitis Patients

Attributes	Value
Class	die (1), live (2)
Sex	male (1), female (2)
Age	numerical value
Anorexia	no (1), yes (2)
Liver Big	no (1), yes (2)
Fatigue	no (1), yes (2)
Malaise	no (1), yes (2)
Steroid	no (1), yes (2)
Antivirals	no (1), yes (2)
Spleen Palpable	no (1), yes (2)
Spiders	no (1), yes (2)
Liver Firm	no (1), yes (2)
Bilirubin	0.39,0.80,1.20,2.00,3.00,4.00
Alk Phosphate	33, 80, 120, 160, 200, 250
Ascites	no (1), yes (2)
Varices	no (1), yes (2)
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
SGOT	13, 100, 200, 300, 400, 500
Histology	no (1), yes (2)

3. Naive Bayes Classification

Naive Bayes[27-30] implements probabilistic naive bayes classifier. Naive means conditional independence among attributers of features. The "naive" assumption greatly reduces computation complexity to a simple multiplication of probabilities. The Naive bayes handles numeric attributes using supervised discretization and uses kernel density estimators that will improve the performance. It needs only small set of training data to develop accurate parameter estimations because it requires only the calculation of the frequencies of attributes and attribute outcome pairs in the training data set[30,31].

4. Data Set

The dataset used in this model collected from UCI machine learning repository[32] should be more precise and accurate in order to improve the predictive accuracy of data mining tasks. The data set may have missing (or) irrelevant attributes and these are to be handled efficiently by the data mining process.

4.1 Attribute Identification

The hepatitis datasetwhich consists of 155 instances and 19 attributes with the class stating the life prognosis yes (or) no. The dataset consist of 6 multi-valued attributes and 14 nominal attributes shown in Table 1.

5. Methodology of Proposed Systems

The research process has two stages. Firstly genetics algorithm based CFS applied to the hepatitis dataset which was reduced to 9 from 19 and the Naive Bayes algorithm applied for classification. The second stage of this work used PSO search based CFS applied to same data set which was reduced to 10 from 19 and then applied Naive Bayes classification algorithm for better prediction. The architectural diagram of proposed research process is shown in Figure 1 and Figure 2.

6. Experimental Results

The experimental results illustrate the various measures that are used to evaluate the model for classification and prediction. In this work the accuracy, sensitivity, specificity, precision and kappa statistics are elaborated.

Table 2. Different Outcome of Two Class Prediction

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

Table 3. Accuracy, Sensitivity, Specificity by class: Before Feature Selection

Classifiers	Accuracy	Sensitivity	Specificity	Precision	Time
J48	83%	83%	94%	82%	0.03
Sequential Minimal Optimization	85%	85%	92%	84%	0.08
Multi Layer Perceptron	80%	80%	86%	80%	17.94
Naive Bayes	84%	84%	89%	85%	0.01
Radial Basis Function	85%	85%	93%	85%	0.03

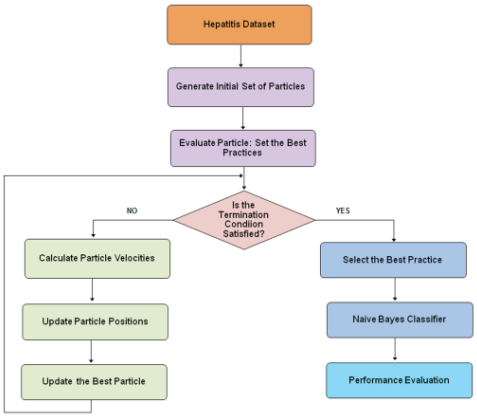

Figure 2. Proposed Methodology for PSO based CFS and Naive Bayes Classification.

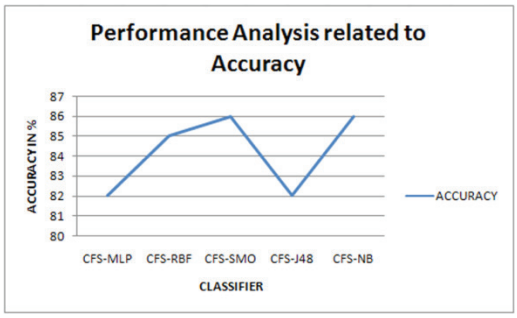

Figure 3. Performance related to accuracy based on Genetic search.

6.1 Accuracy, Sensitivity, Specificity and Precision

A single prediction has the four different possible outcomes shown in Table 2. The following equation that are used for measuring the classification Accuracy Equation (4), Sensitivity Equation (5), Specificity Equation (6) and Precision Equation (7).

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

$$\text{Sensitivity} = TP/(TP + FN) \quad (5)$$

$$\text{Specificity} = TN/(TN + FP) \quad (6)$$

$$\text{Precision} = TP/(TP + FP) \quad (7)$$

Table 3 shows the Accuracy, Time, Precision, Sensitivity and Specificity that are calculated for the top classification algorithms that were applied to the hepatitis data set from the previous studies[31,33-35].

Table 4 illustrates the Accuracy, Time, Precision, Sensitivity and Specificity that are calculated for the PSO search based on correlation based Feature section with

Table 4. Accuracy, sensitivity, specificity by Class: After Feature Selection based PSO search

Classifiers	Accuracy	Sensitivity	Specificity	Precision	Time
PSO based CFS and J48	83%	83%	95%	81%	0.02
PSO based CFS and SMO	83%	83%	94%	81%	0.03
PSO based CFS and MLP	83%	83%	88%	84%	0.28
PSO based CFS and NB	84%	84%	88%	85%	0.01
PSO based CFS and RBF	84%	84%	92%	84%	0.02

Table 5. Accuracy, sensitivity, specificity by Class: After Feature Selection based on Genetic Search

Classifiers	Accuracy	Sensitivity	Specificity	Precision	Time
Genetic based CFS and J48	82%	82%	94%	80%	0.07
Genetic based CFS and SMO	86%	86%	90%	87%	0.02
Genetic based CFS and MLP	82%	82%	91%	81%	0.35
Genetic based CFS and NB	86%	86%	90%	87%	0.01
Genetic based CFS and RBF	85%	85%	92%	84%	0.02

Table 6. Before Feature Selection based on Naive Bayes

a	b	Classified as
22	10	a = DIE
14	109	b = LIVE

Table 7. After Feature Selection for PSO search based CFS and Naive Bayes

a	b	Classified as
22	10	a = DIE
14	109	b = LIVE

Table 8. After Feature Selection for Genetic search based CFS and Naive Bayes.

a	b	Classified as
22	9	a = DIE
12	111	b = LIVE

multilayer perceptron, radial basis function, sequential minimal optimization, J48 and naive bayes.

Table 5 shows the Accuracy, Time, Precision, Sensitivity and Specificity that are calculated for the genetic search based on correlation based Feature section with multilayer perceptron, radial basis function, sequential minimal optimization, J48 and naive bayes.

The model was evaluated based on classification accuracy, sensitivity and specificity and precision. This work has used PSO search based CFS and Naive Bayes algorithm and also genetic search based CFS and naive bayes algorithm. The model has achieved 84% of predictive accuracy based on PSO based search and 86% for genetic search.

6.2 Kappa Statistics

Kappa statistics was introduced by Cohen which measures the agreement of prediction with the true class and a 1 signifies complete agreement[36]. The kappa value for

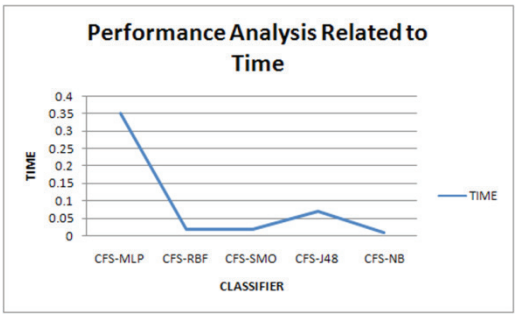

Figure 4. Performance related to time based on Genetic search.

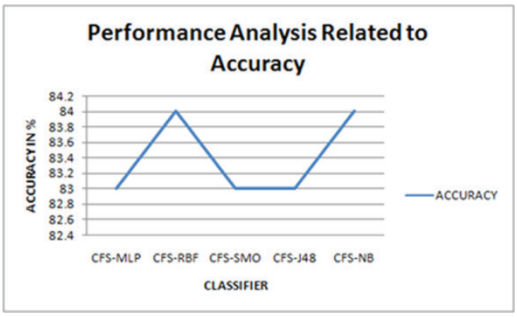

Figure 5. Performance related to accuracy based on PSO search.

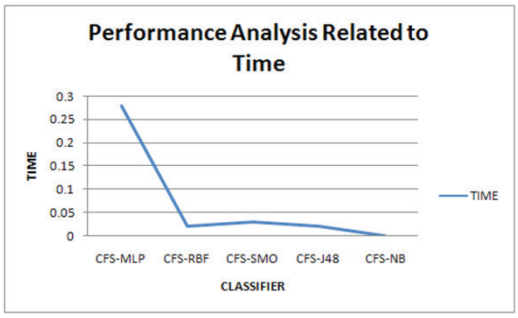

Figure 6. Performance Related to Time based on PSO search.

PSO search based CFS and Naive Bayes is 0.5483 and for genetic search based CFS and Naive Bayes is 0.6004.

6.3 10-Fold Cross-Validation

The classification algorithm is trained and tested in 10 times. The cross validation divides the data into 10 subgroups and each subgroup is tested through classification rule constructed from the remaining 9 groups. Ten different test results are obtained for each train–test configuration and the average result provides the test accuracy of the algorithm.

6.4 Confusion Matrix

The confusion matrix[7] illustrates how many instances have been assigned to each class and the elementsof the matrix illustrates the number of test examples whose actual class is the row and whose predicted class is the column. Tables 6, 7 and 8 illustrate the confusion matrix that is calculated for Naive Bayes and the genetic search based CFS-NB and PSO based CFS-NB algorithms.

6.5 Graph Results

Figure 3 show the accuracy of various classification algorithms that was achieved through genetic search based CFS.

Figure 4 shows the time that was calculated over various classification algorithms based on Genetic search based CFS.

Figure 5 shows the accuracy of various classification algorithms that was achieved through PSO search based CFS.

Figure 6 shows the time that was calculated over various classification algorithms based on PSO search based CFS.

7. Conclusion

In this work, an enhanced method was developed for hepatitis diagnosis. The results shows that PSO search based CFS-NB achieved the same classification accuracies for a reduced feature subset that contained ten features. The genetic search based CFS-NB was produced better clas-

sification accuracy for reduced subset of nine features. The comparative study was conducted on the hepatitis data based on PSO and genetic search based on CFS with other classification algorithms like J48, Radial basis function, Multilayer perceptron and Sequential Minimal Optimization. The experimental result clearly illustrates that the genetic search based CFS and naive bayes performance was better compared with other classification algorithms in terms of time and accuracy.

8. References

1. Tan KC, Toeh EJ, Yua Q, Goh KC. A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications. 2009; 36:8616–30.
2. Yoon H, Park C-S, Kim JS, Baek JG. Algorithm learning based neural network integrating feature selection and classification. Expert Systems with Applications. 2013; 40:231–41.
3. Chandrasekar G, Sahin F. A survey on feature selection methods. Computers and Electrical Engineering. 2014; 40:16–24.
4. Chandra B, Gupta M. An efficient statistical feature selection Approach for classification of gene expression data. Journal of Bio medical Informatics. 2011; 44:529–35.
5. Liu H, Yu L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering. 2005 Apr; 17(4).
6. Purusothaman G, Krishnakumari P. A Survey of Data Mining Techniques on Risk Prediction: Heart Disease. Indian Journal of Science and Technology. 2015 Jun; 8(12).
7. Suganya P, Sumathi CP. A Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease. Indian Journal of Science and Technology. 2015 Jul; 8(14).
8. Kalaiselvi C, Nasira GM. Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques. Indian Journal of Science and Technology. 2015 Jul; 8(14).
9. Hall MA, Smith LA. Feature Selection for Machine Learning: Comparing a Correlation based filterapproach to the wrapper. FCAIRS Conference; 1999.
10. Hall M. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the Seventeenth International Conference on Machine Learning; 2000. p. 359–66.
11. Lu X, Peng X, Deng Y, Feng B, Liu Ping, Liao B. A Novel Feature Selection Method Based on Correlation-Based Feature Selection in Cancer Recognition. Journal of computational and Theoretical nanoscience. 2014 Feb; 11(2):427–33.
12. Eberhart RC, Kennedy J. A new optimizer using particle swarm theory. In Proceedings of the sixth international symposium on micro machine and human science. 1995; 1:39–43.
13. Shi Y, Eberhart RC. Empirical study of particle swarm optimization. IEEE Proceedings of the 1999 Congress on Evolutionary Computation; 1999. p. 3.
14. Eberhart RC, Shi Y. Particle swarm optimization: developments, applications and resources. IEEE Proceedings of the 2001 Congress on in Evolutionary Computation. 2001; 1: 81–6.
15. Vijayarani S, Sudha S. An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples. Indian Journal of Science and Technology. 2015 Aug; 8(17).
16. Caballero-Ruiz E, García-Saez G, Balsells M, Pons B, Morillo M, Gomez EJ, Hernando ME. Automatic Blood Glucose Classification for Gestational Diabetes with Feature Selection: Decision Trees vs. Neural Networks. XIII Mediterranean Conference on Medical and Biological Engineering and Computing, 2013 IFMBE Proceedings. 2014; 41:1370–3.
17. Xu J, Sun L, Gao Y, Xu T. An ensemble feature selection technique for cancer recognition. Bio-Medical Materials and Engineering. 2014; 24:1001–8.
18. Sridevi T, Murugan A. A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis. International Journal of Computer Applications. 2014 Feb; 88(11).
19. Wang J, Zhou S, Yi Y, Kong J. An Improved Feature Selection Based on Effective Range for Classification. The Scientific World Journal; 2014. p. 1–8.
20. Zang Y, Yang A, Xiong C, Wang T, Zhang Z. Feature Selection for Data Envelopment Analysis. Knowledge Based Systems. 2014 Jul; 64:70–80.
21. Ding S, Li Y, Shi Z, Yan S. A Protein Structural Classes Prediction Method based on Predicted Secondary Structure and PST-BLAST Profile. Biochimie. 2014 Feb; 97:60-5.
22. Riccardo P, Kennedy J, Blackwell T. Particle swarm optimization. Swarm intelligence. 2007; 1(1):33–57.
23. Kennedy J. Particle swarm optimization. Encyclopedia of Machine Learning. Springer US; 2010. p. 760–6.
24. Mandal M, Mukhopadhyay A. A novel PSO-based graph-theoretic approach for identifying most relevant and non-redundant gene markers from gene expression data. International Journal of Parallel, Emergent and Distributed Systems; 2014. p. 1–18.
25. Ji Z, Wang B. Identifying potential clinical syndromes of Hepatocellular carcinoma Using PSO-based hierarchical feature selection algorithm. Bio Med research international; 2014. p. 1–12.
26. Yasodha P, Ananthanarayanan NR. Analyzing Big Data to Build Knowledge Based System for Early Detection of

Ovarian Cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14).
27. Dumitru D. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. Annals of University of Craiova Math Comp Sci Ser. 2009; 36(2):92–6.
28. Nagarajan S, Chandrasekaran RN. Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques. Indian Journal of Science and Technology. 2015 Apr; 8(8):771–6.
29. Anuradha C, Velmurugan T. A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. Indian Journal of Science and Technology. 2015 Jul; 8(15).
30. Leung KS, Lee KH, Wang JF, Ng EY, Chan HL, Tsui SK, Mok TS, Tse pC, Sung JJ. Data Mining on DNA Sequences of Hepatitis B Virus. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2011 Mar-Apr; 8(2):428–40.
31. Karthikeyan T, Thangaraju P. Analysis of Classification Algorithms Applied to Hepatitis Patients. International Journal of Computer Applications. 2013 Jan; 62(5):25–30.
32. Available from: http://archieve.ics.uci.edu/ml/datasets/hepatitis/ Accessed on: 27.05.2015.
33. Karthikeyan T, Thangaraju P. PCA-NB Algorithm to Enhance the Predictive Accuracy. International Journal of Engineering and Technology. 2014 Feb-Mar; 6(1):381–7.
34. Karthikeyan T, Thangaraju P. A Combined Approach of CFS and Naive bayes Algorithm to Enhance the Prediction Accuracy. IEEE International Conference on Control Instrumentation Communication and Computational Technologies (ICCICCT 2014); 2014 Jul.
35. Karthikeyan T, Thangaraju P. Best First and Greedy Search based CFS-Naive Bayes Classification Algorithms for Hepatitis Diagnosis, Biosciences and Biotechnology Research Asia. 2015 Apr; 12(1):983–90.
36. Written IH, Frank E. Data mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers; 2011.