BIOSCIENCES BIOTECHNOLOGY RESEARCH ASIA, April 2015. Vol. **12**(1), 983-990

Best First and Greedy Search Based CFS- Naïve Bayes Classification Algorithms for Hepatitis Diagnosis

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

[1]Department of Computer Science, PSG College of Arts and Science, Coimbatore, India.
[2]Department of Computer Applications, Bishop Heber College, Tiruchirappalli, India.

DOI: http://dx.doi.org/10.13005/bbra/1749

(Received: 06 February 2015; accepted: 25 March 2015)

The main purpose of this paper is to deal with classification algorithm with feature selection is used to improve the prediction accuracy in the medical data. This paper applies best first search and greedy search as a searching methods and feature evaluator used as CFS. Naive Bayes classification algorithm is used for hepatitis patients' dataset. It analyzes the data set taken from the UC Irvine machine learning repository. The result of the classification model is time and improved classification accuracy. Finally, it concludes that the proposed methodology performance is better than other classification algorithms.

Key words: Classification, Correlation Based Feature Selection, Feature Selection, Hepatitis, Naïve Bayes.

Hepatitis means injury to the liver with inflammation of the liver cells. The liver is the largest gland in the human body. It weighs approximately 3 lb (1.36 kg). It is reddish brown in color and is divided into four lobes of different sizes and lengths. It is also the largest internal organ. It is below the diaphragm on the right in the thoracic region of the abdomen. Blood reaches the liver through the hepatic artery and the portal vein. The portal vein carries blood containing digested food from the small intestine, while the hepatic artery carries oxygen-rich blood from the aorta.

The liver is made up of thousands of lobules; each lobule consists of many hepatic cells. Hepatic cells are the basic metabolic cells of the river. The liver has a wide range of functions, including: Detoxification, Stores vitamins A, D, K and B12 Protein synthesis. The production of biochemical needed for digestion, such as bile, Maintains proper levels of glucose in the blood, produces 80% of your body's cholesterol, the storage glycogen, decomposing red blood cells, synthesizing plasma protein, the production of hormones and produces urea.

Most liver damage is caused by 3 hepatitis viruses, called hepatitis A, B and C. However, hepatitis can also be caused by alcohol and some other toxins and infections, as well as from our own autoimmune process. About 250 million people globally are thought to be affected by hepatitis C, while 300 million people are thought to be carriers of hepatitis B. Not all forms of hepatitis are infectious. Alcohol, medicines, and chemical may be bad for the liver and cause inflammation.

A person may have a genetic problem, a metabolic disorder, or an immune related injury. Obesity can be a cause of liver damage which can lead to inflammation. These are known as non-infectious, because they cannot spread form person-to-person. Life prognosis of hepatitis is a challenging task in early stage due to various

* To whom all correspondence should be addressed
E-mail: thangarajubhc@yahoo.co.in

interdependent features. A model can be developed which can used in life prognosis of hepatitis diseases. Data mining techniques have been extensively used in bioinformatics to analyze biomedical data. Data mining algorithms can be used efficiently in prediction and classification of inter-related data. The objective of this analysis is classify and improving the accuracy of hepatitis data base.

Feature Selection[1-2] is a technique which is used to reduce the dimensionality of data or eliminate the irrelevant features and to improve 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[3-4].

The objective of this study is to predict the life expectancy for patients with hepatitis based on a hepatitis data and improve the classification accuracy. The proposed work is using Naive Bayes algorithm to get the accuracy of the classification and prediction. In order to increase its accuracy Correlation Based Feature Selection (CFS), best first algorithm and greedy approach of feature selection is being used. This is to make sure the noisy or irrelevant features are removed. Then compare the accuracy of prediction by using Naive bayes and other classification algorithms like J48, Multi layer Perceptron (MLP), Radial Basis Function (RBF).

This paper is organized as follows. Part 2 deals with related work. Part 3 deals with the concept of CFS and Best First Search based CFS and Naïve Bayes Algorithm (BFSCFS-NB) and greedy search based CFS and Naïve Bayes algorithm (GSCFS-NB). Part 4 elaborates the naive bayes classification algorithm. Part 5 discusses the data set descriptions. Part 6 deals with the proposed methodology and part 7 illustrate the performance evaluation.

Related Works

Lu, Xinguo *et al.*,[5] has 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.

E.Caballero-Ruiz *et al.*[6] dealt with classification with automatic blood glucose data from patients glucose meter for the development of decision support systems for gestational

Table 1. Attribute details of the hepatitis patients

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

Table 2. Different outcome of two class prediction

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

diabetes. They used feature selection methods and neural networks and decision trees classification algorithm for their research.

Jiucheng Xu *et al.*,[7] 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.

T.Sridevi and A.Murugan[8] 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.

Correlation based feature selection

CFS[9-12] is one of well-known techniques to rank the relevance of features by measuring correlation between features and classes and between features and other features. Given number

Table 3. Detailed accuracy by class: before feature selection

Classification Algorithms	Accuracy	Time	Precision	Sensitivity	Specificity
Naive Bayes	84%	0.0	85%	84%	89%
J48	83%	0.03	82%	83%	94%
Multi Layer Perceptron	80%	17.94	80%	80%	86%
SMO	83%	0.08	83%	84%	91%
RBF	83%	0.03	84%	84%	92%

Table 4. Detailed accuracy by class: after feature selection

Classification Algorithms	Accuracy	Time	Precision	Sensitivity	Specificity
BFS based CFS-MLP	84%	0.45	84%	84%	91%
BFS based CFS-RBF	86%	0.17	85%	86%	93%
BFS based CFS-SMO	83%	0.06	81%	83%	94%
BFS based CFS-J48	81%	0.05	79%	81%	92%
BFS based CFS-NB	88%	0.01	87%	87%	91%

Table 5. Detailed accuracy by class: after feature selection

Classification Algorithms	Accuracy	Time	Precision	Sensitivity	Specificity
GS based CFS-MLP	84%	0.25	84%	84%	91%
GS based CFS-RBF	86%	0.2	85%	86%	93%
GS based CFS-SMO	83%	0.02	81%	83%	94%
GS based CFS-J48	81%	0.01	79%	81%	92%
GS based CFS-NB	88%	0.01	87%	87%	91%

Table 6. Before feature selection

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

Table 7. After feature selection

a	b	Classified as
23	9	a = DIE
10	113	b = LIVE

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 *Ms* 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 research, we used best first search (BFS) and greedy hill climbing search algorithms for the best results[13-14].

GSCFS-NB Algorithm

Searching the space of feature subsets within reasonable time constraints is necessary if a feature selection algorithm is to operate on data with a large number of features. One simple search strategy, called greedy hill climbing, considers local changes to the current feature subset. Frequently, local change is that is the addition or deletion of a single feature from the subset. When the algorithm considers only additions to the feature subset it is known as forward selection and the deletions is known as backward elimination method[6,13,14].

An alternative approach, called stepwise bi-directional search, uses both addition and deletion. It encompasses each of these variations, the search method may consider all possible local changes to the current subset and then choose the best, or the first change that improves the merit of the current feature subset. In both cases, once a change is identified then it is never reconsidered. The first half(step 1 to 7) of the algorithm is used to select the subset using Genetic Search and then the second half(8 to 12) of the algorithm is for classification using Naïve Bayes. The GSCFS-NB classification algorithm is given below.

Step 1 Let s ← start state.
Step 2 Enlarge s by making each possible local change.
Step 3 Evaluate each child t of s.
Step 4 Let s' ← child t with highest evaluation e (t).
Step 5 If e(s') ≥ e(s) then s ←s', go to 2.
Step 6 Return s.
Step 7 Obtain the new data set.
Step 8 Construct both training and test data discrete.
Step 9 Estimate the prior probabilities $P(C_j)$, j=1,... k from the training data, where k is the number of classes.
Step 10 Estimate the conditional probabilities $P(A_i = a_{\ell} / C_j)$, i= 1,....,D, j=1,....,k, $-\ell$= 1,....,d from the training data, where D is the number of features, d is the number of discretization level.
Step 12 Estimate the posterior probabilities $P(C_j / A)$ for each test example x represented by a feature vector A.
Step 13 Assign x to the class C^* such that C^*= arg max$_{j=1,2}$ $P(C_j/A)$.

BFSCFS-NB Algorithm

The Best first search is an important AI search strategy that allows backtracking along the search path[13, 14]. Like greedy hill climbing, best first moves through the search space by making local changes to the current feature subset. However,

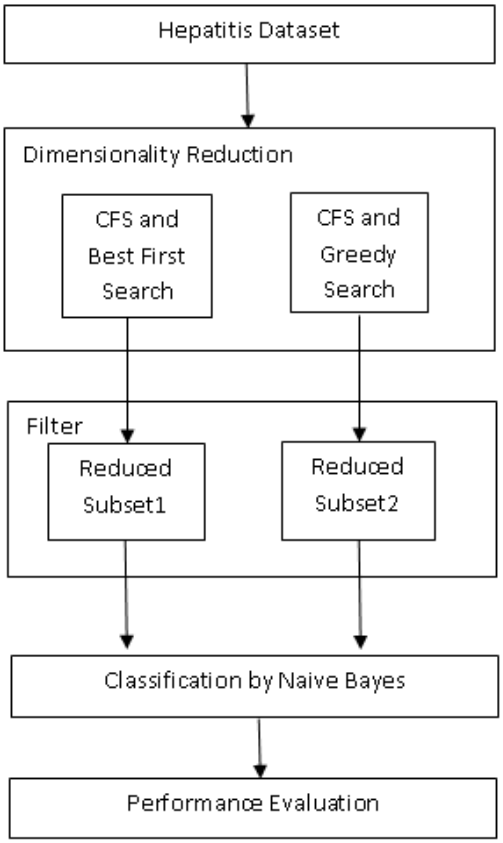

Fig. 1. Proposed methodology

unlike hill climbing method, suppose path being explored begins to look less promising, the best first search method can back-track to a more promising previous subset and continue the search from there. A best first search will explore the entire search space for specified time, so it is common to use a stopping criterion. Normally this involves limiting the number of fully expanded subset that result in no improvement. The first half(step 1 to 8) of the algorithm is used to select the subset using Best First Search and then the second half(step 9 to 13) of the algorithm is for classification using Naïve Bayes. The following shows the BFSCFS-NB classification algorithm.

Step 1 To start with OPEN list containing the start state, the CLOSED list empty and BEST ← start state.

Step 2 Let assign s = arg max e(x).

Step 3 Eliminate s from OPEN and add to CLOSED.

Step 4 If e(s) ≥ e (BEST), then BEST ← s.

Step 5 For every child t of s that is not in the OPEN or CLOSED list, evaluate and add to OPEN.

Step 6 If BEST changed in the last set of expansions then go to 2.

Step 7 Return BEST.

Step 8 Obtain the new data set.

Step 9 Construct both training and test data discrete.

Step 10 Estimate the prior probabilities $P(C_j)$, j=1,... k from the training data, where k is the number of classes.

Step 11 Estimate the conditional probabilities $P(A_i = a_f / C_j)$, i= 1,....,D, j=1,....,k, f= 1,....,d from the training data, where D is the number of features, d is the number of discretization level.

Step 12 Estimate the posterior probabilities $P(C_j / A)$ for each test example x represented by a feature vector A.

Step 13 Assign x to the class C^* such that $C^* = \arg\max_{j=1,2} P(C_j / A)$.

Naïve Bayes Classifier

A Naive Bayesian[15-17] classifier based on Bayes theorem is a probabilistic statistical classifier. Here, the term "naive" indicates conditional independence among features or attributes. The "naive" assumption greatly reduces computation complexity to a simple multiplication

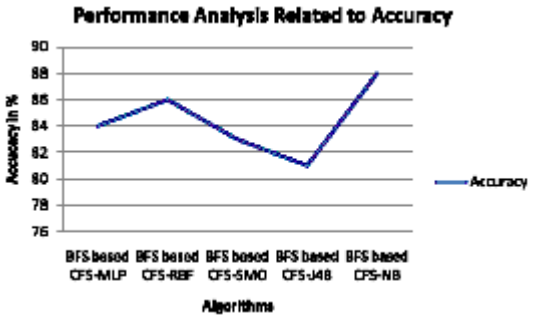

Fig. 2. Performance related to accuracy

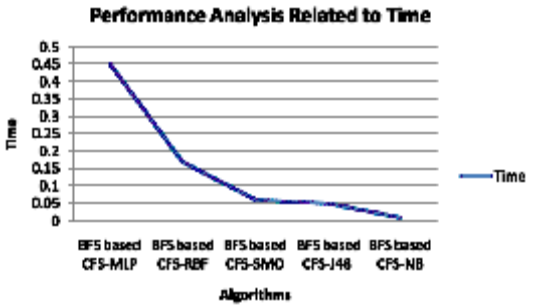

Fig. 3. Performance related to time

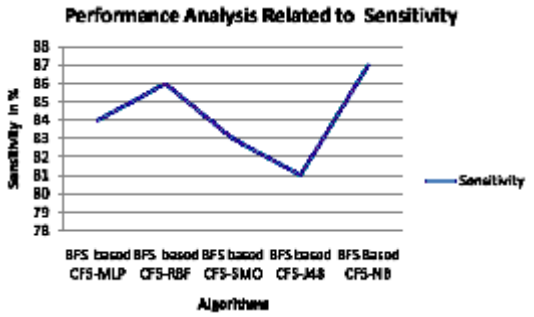

Fig. 4. Performance related to accuracy

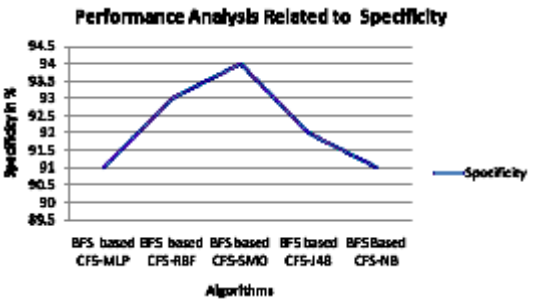

Fig. 5. Performance related to time

of probabilities. The prime advantage of the Naive Bayesian classifier is its swiftness of use and this swiftness occurs because it is the simplest algorithm among classification algorithms. Because it is simple, it can eventually handle a data set with many attributes.

The naive Bayesian classifier 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[18]. In this paper, Naive Bayes algorithm is used as a classification algorithm.

Data set

The dataset used in this model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. The dataset set is collected may have missing (or) irrelevant attributes. These are to be handled efficiently to obtain the optimal outcome from the data mining process.

Attribute Identification

Dataset collected from UC Irvine machine learning repository[19] which consists of 155 instances and 19 attributes with the class stating the life prognosis yes (or) no. The dataset consist of 14 nominal attributes and 6 multi-valued attributes shown in Table 1.

Methodology of Proposed System

The new approach is incorporated in two stages. Firstly all the number of features of the hepatitis disease dataset was reduced to 10 from 19 by CFS Evaluator based on best first and greedy search. Then, hepatitis disease dataset is classified by using Naive Bayes classification Algorithm. The block diagram of proposed methodology is shown in Fig. 1.

Performance Evaluation

It needs a measure for evaluating performance which has to be introduced and this measure in the literature is accuracy defined as correct classified instances divided by the total number of instances.

Accuracy, Sensitivity, Specificity and Precision

A single prediction has the four different possible outcomes shown in Table II for The true positives (TP) and true negatives (TN) are correct classifications. In this work false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually yes. We use the following equation to measure the accuracy Eq. (2), sensitivity Eq. (3), specificity Eq. (4), Precision Eq. (5)

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

$$\text{Sensitivity} = TP/(TP + FN) \quad ...(3)$$

$$\text{Specificity} = TN/(TN + FP) \quad ...(4)$$

$$\text{Precision} = TP/(TP + FP) \quad ...(5)$$

The accuracy, Time, Precision, Sensitivity and Specificity for Naive Bayes, J48, Multilayer Perceptron, SMO and RBF from the previous studies[20-21] are shown in Table III.

The accuracy, Time, Precision, Sensitivity and Specificity for BFS based CFS-MLP, BFS based CFS-RBF, BFS based CFS-SMO, BFS based CFS-J48 and BFS based CFS-NB are list out in Table IV.

Based on classification accuracy, sensitivity and specificity the models were evaluated. We have applied CFS with Best first search and naïve bayes in our proposed method. Using this model a prediction accuracy of 88% is achieved.

The accuracy, Time, Precision, Sensitivity and Specificity for GS based CFS-MLP, GS based CFS-RBF, GS based CFS-SMO, GS based CFS-J48 and GS based CFS-NB are mentioned in Table V.

K-Fold Cross-Validation

The k-fold cross-validation method has been used for best performance in this work. The classification algorithm is trained and tested k time. In its most elementary form, cross validation divides the data into k subgroups and each subgroup is tested via classification rule constructed from the remaining (k -1) groups. Thus the k different test results are obtained for each train–test configuration. The average result gives the test accuracy of the algorithm. It uses tenfold cross-validation in this work.

Kappa Statistics

The kappa parameter measures pair wise agreement between two different observers, corrected for an expected chance agreement. For instance, if the value is one, it means that there is a complete agreement between the classifier and real world value. Kappa value is calculated using

following equation

$$K = [P(A) - P(E)/[1 - P(E)] \quad ...(6)$$

$$P(A) = (TP + TN) / N \quad ...(7)$$

$$P(E) = [(TP + FN) * (TP + FP) * (TN + FN)]N^2 \quad ...(8)$$

Where N is the total number of instances used. P (A) is the percentage of agreement between the classifier and underlying truth calculated by Eq. (7). P (E) is the chance of agreement calculated by Eq. (8). In this study the kappa value is 0.6302 for BFSCFS-NB and GSCFS- NB which is calculated by Eq. (6).

Confusion Matrix

A confusion matrix is calculated for Naive Bayes, BFSCFS-NB and GSCFS-NB classifiers based on best first and greedy search to interpret the results. The confusion matrix is shown in Tables VI and VII.

Graph Results

Fig. 2. Shows the performance analysis related to accuracy of various algorithms based on CFS and Best First search.

Fig. 3. Shows the performance analysis related to time over various algorithms based on CFS and best first search

Fig. 4. Shows performance analysis related to accuracy of various algorithms based on CFS and Greedy search.

Fig. 5. Shows the performance analysis related to time of various algorithms based on CFS and Greedy search.

CONCLUSION

In this proposed work an enhanced medical diagnostic method for addressing hepatitis diagnosis problem is developed. Experiment results on various portions of the hepatitis dataset proved that the new approach performs better in distinguishing the live from the dead one. It is observed that BFSCFS-NB and GSCFS-NB achieved the best classification accuracies for a reduced feature subset that contained ten features. Meanwhile, comparative study is conducted on the methods such as MLP, SMO, J48, and RBF. The experimental result shows that the BFSCFS-NB and GSCFS-NB performed advantageously over the other methods in terms of the classification accuracy and time. We hope the results demonstrated by the proposed algorithms can ensure that the physicians make accurate diagnostic decision.

REFERENCES

1. Yoon, H., Park, C. S., Kim, J. S., Baek, J. G. Algorithm learning based neural network integrating feature selection and classification. *Expert Systems with Applications*, 2013; **40**(1): 231-241.
2. Tan, K. C., Teoh, E. J., Yu, Q., Goh, K. C. A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, 2009; **36**(4): 8616-8630.
3. Liu, H., Yu, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2005; **17**(4): 491-502.
4. Chandra, B., Gupta, M. An efficient statistical feature selection approach for classification of gene expression data. *Journal of biomedical informatics*, 2011; **44**(4): 529-535.
5. Lu, X., Peng, X., Deng, Y., Feng, B., Liu, P., Liao, B. A novel feature selection method based on correlation-based feature selection in cancer recognition. *Journal of Computational and Theoretical Nanoscience*, 2014; **11**(2): 427-433.
6. Caballero-Ruiz, E., García-Sáez, G., Rigla, M., Balsells, M., Pons, B., Morillo, M., Hernando, M. E. Automatic Blood Glucose Classification for Gestational Diabetes with Feature Selection: Decision Trees vs. Neural Networks. In: XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013; *Springer International Publishing*, 2014; **41**: 1370-1373.
7. Xu, J., Sun, L., Gao, Y., Xu, T. An ensemble feature selection technique for cancer recognition. *Bio-Medical Materials and Engineering*, 2014; **24**(1): 1001-1008.
8. Sridevi, T., Murugan, A. A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis. *International Journal of Computer Applications*, 2014; **88**(11): 28-33.
9. Hall, M. A., Smith, L. A. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In: FLAIRS conference, May 1999; 235-239.
10. Wang, J., Zhou, S., Yi, Y., Kong, J. An Improved Feature Selection Based on Effective Range for Classification. *The Scientific World Journal*, 2014.
11. Zhang, Y., Yang, A., Xiong, C., Wang, T., Zhang, Z. Feature selection using data envelopment

analysis. *Knowledge-Based Systems*, 2014; **64**: 70-80.

12. Santana, L. E. A. D. S., Canuto, A. M. Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications*, 2014; **41**(4): 1622-1631.
13. Hall. M. A. Correlation based feature selection for machine learning, Doctoral dissertation, University of Waikato, *Department of Computer Science*, 1999.
14. Hall. M, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the Seventeenth International Conference on Machine Learning, 2000; 359-366.
15. Ding. S., *et al*., A Protein Structural Classes Prediction Method based on Predicted Secondary Structure and PST-BLAST Profile, *Biochimie*, 2014; **97**: 60-65.
16. Feng, P. M., Ding, H., Chen, W., Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. Computational and mathematical methods in medicine, 2013.
17. Leung, K. S., Lee, K. H., Wang, J. F., Ng, E. Y., Chan, H. L., Tsui, S. K., Sung, J. Y. Data mining on dna sequences of hepatitis b virus. *IEEE/ ACM Transactions on Computational Biology and Bioinformatics*, 2011; **8**(2): 428-440.
18. Dumitru, D. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. *Annals of the University of Craiova-Mathematics and Computer Science Series*, 2009; **36**(2): 92-96.
19. http://archieve.ics.uci.edu/ml/datasets/hepatitis/
20. Karthikeyan. T., Thangaraju, P. Analysis of Classification Algorithms Applied to Hepatitis Patients, *International Journal of Computer Applications*, 2013; **62**(5): 25-30.
21. Karthikeyan. T., Thangaraju, P. PCA-NB Algorithm to Enhance the Predictive Accuracy, *International Journal of Engineering and Technology*, 2014; **6**(1): 381-387.