PCA-NB Algorithm to Enhance the Predictive Accuracy

T.Karthikeyan[1], P.Thangaraju[2]
[1]Associate Professor, Dept. of Computer Science, P.S.G Arts and Science College, Coimbatore, India
[2] Research Scholar, Bharathiar University, Asst. Professor, Dept. of Comp. Applications, Bishop Heber College, Tiruchirappalli, India
[1] *t.karthikeyan.gasc@gmail.com*
[2] *thangarajubhc@yahoo.co.in*

***Abstract*- This paper mainly deals with feature extraction algorithm used to improve the predicted accuracy of the classification. This paper applies with Principal Component analysis as a feature evaluator and ranker for searching method. Naive Bayes algorithm is used as a classification algorithm. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time. Finally, it concludes that the proposed PCA-NB algorithm performance is better than other classification techniques for hepatitis patients.**

Keyword- Feature Extraction, Classification, Principal Component Analysis, Naive Bayes

I. INTRODUCTION

In data mining and in image processing, feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction [1].

If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Feature extraction can be applied in many data mining applications to improve the predictive accuracy.

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. We are going to use Naive Bayes algorithm to get the accuracy of the classification and prediction. In order to increase its accuracy Principal Component Analysis [2] of feature reduction is being used. This is to make sure the noisy or irrelevance feature should be taken care of. 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. The section 2 deals with related work. Section 3 deals with the concept of feature extraction and principal component analysis. Section 4 elaborates with the naive bayes classification algorithm. Section 5 discusses with the data set descriptions. Section 6 deal with the proposed methodology and section 7 illustrates the performance evaluation.

II. RELATED WORK

Many feature extraction methods are used to deal with the diagnosis of medical diagnosis problem, and most of them have achieved better classification accuracies.

Kemal Polat et.al. Used an artificial immune recognition system and principal component analysis (PCA) via 10-fold cross-validation was used for classification [3]. Tahseen a jilani et.al. used PCA-ANN based classification algorithm for hepatitis disease diagnosis[4]. Heng lian used principal component analysis and one class support vector machinesl for image retrieval[5]. Hlip-ling chen et at. Used hybrid prediction model which integrates a local discriminant analysis and support vector machines for hepatitis disease diagnosis [6].

Yilmaz Kaya and Murat Uyar developed a hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease [7]. Huda yasin et al. Uses PCA as a feature extraction tool and regression analysis and achieves 89% classification accuracy [8]. Javad salami satakhti et. al. applied support vector machine and simulated anneling for hepatitis disease diagnosis[9].

In order to improve the classification accuracy, PCA feature extraction method is used. The objective of the proposed method is to explore the performance of hepatitis diagnosis using an algorithm that integrates PCA with Naive Bayes.

The proposed method (PCA-NB) is firstly to use PCA in reducing the dimension of the hepatitis dataset, and then the obtained reduced feature subset is served as the input into the designed NB classifier. The effectiveness of PCA-NB is examined in terms of classification accuracies, sensitivity and specificity, precision.

Further, the superior classification capability of the proposed method can be observed by comparing the results with those using MLP based on PCA (PCA-MLP), RBF based on PCA(PCA-RBF), J48 based PCA(J48-PCA), Random Forest based PCA(PCA-RF) and the standard NB. Experimental results have shown that PCA-NB outperforms the other methods significantly and has achieved the best predicative classification accuracy with the reduced feature subset.

III. PCA FOR FEATURE EXTRACTION AND REDUCTION

PCA is a statistical technique used for extracting information from a multi-variety dataset. This process is performed via having principal components of original variables with linear combinations identified. While the original dataset with the maximum variability is represented with the first principal component, the dataset from the remaining with the maximum variability is represented with the second principal component [1, 3]. The process goes on consecutively as such, with the dataset from the remaining with the maximum variability being represented with the next principal component. While *m* represents the number of all principal components, and *p* represents the number of the significant principal components among all principal components, *p* may be defined as the number of those principal components of the *m* dimensional dataset with the highest variance values. It is clear therein that $p \leq m$. Therefore, PCA may be defined as a data-reducing technique. In other words, PCA is a technique used for producing the lower-dimensional version of the original dataset.

IV. NAIVE BAYES CLASSIFIER

A Naive Bayesian 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 of probabilities. The major advantage of the Naive Bayesian classifier is its rapidity of use. This rapidity occurs because it is the simplest algorithm among classification algorithms. Because of this simplicity, it can readily handle a data set with many attributes. In addition, 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 [10, 11]. In this paper Naive Bayes algorithm is used as a classification algorithm.

V. DATA SET

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

A. Attribute Identification

Dataset collected from UC Irvine machine learning repository 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 I.

TABLE I
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)

VI. METHODOLOGY OF PROPOSED PCA-NB

The proposed approach consists of two stages. First of all, the number of feature of the hepatitis disease dataset was reduced to 16 from 19 by principal component analysis. Then, hepatitis disease dataset is classified by using Naive Bayes classifier system. The block diagram of proposed methodology is shown in Fig 1.

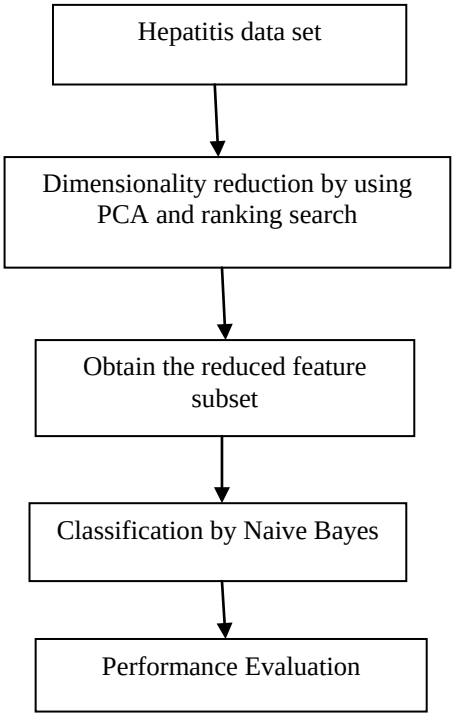

Fig. 1. Proposed methodology

The proposed PCA-NB algorithm is given below:

Step 1: Get the data set.

Step 2: Subtract the mean.

Step 3: Compute the covariance matrix.

Step 4: Compute the eigenvectors and Eigen values of the covariance matrix.

Step 5: Choose components and forming a feature vector.

Step 6: Derive the new data set.

Step7: Make both training and test data discrete;

Step 8: Estimate the prior probabilities P(Cj),j= ,... k from the training data, where k is the number of classes;

Step 9: Estimate the conditional probabilities $P(Ai= a_{\ell} \mid Cj)$,i= 1,....,D,J=1,....,k, ℓ= 1,....,d from the training D is the number of features, d is the number of discretization level;

Step 10: Estimate the posterior probabilities $P(Cj \mid A)$ for each test example x represented by a feature vector A;

Step 11: Assign x to the class C^* such that $C^* = \arg\max_{j=1,2} P(Cj \mid A)$;

VII. PERFORMANCE EVALUATION

Some measure of evaluating performance has to be introduced. One common measure in the literature is accuracy defined as correct classified instances divided by the total number of instances.

A. Accuracy, Specificity, Sensitivity, 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. A 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. In this study we use following equation to measure the accuracy Eq. (1), specificity Eq. (2), sensitivity Eq. (3), Precision Eq. (4).

Accuracy = (TP+TN) / (TP+TN+FP+FN) (1)

Sensitivity= TP / (TP+FN) (2)

Specificity = TN / (TN+FP) (3)

Precision = TP / (TP+FP) (4)

TABLE II
Different outcome of two class prediction

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

The accuracy, Precision, Sensitivity and Specificity for Naive Bayes, J48, Multilayer Perceptron, SMO and RBF are shown in Table III.

TABLE III
Detailed accuracy by class: Before Feature Extraction

Classification Algorithms	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
Naive Bayes	84%	85%	84%	89%
J48	83%	82%	83%	94%
Multi Layer Perceptron	80%	80%	80%	86%
SMO	85%	84%	85%	92%
RBF	85%	85%	85%	93%

The accuracy, Precision, Sensitivity and Specificity for PCA-MLP, PCA-RBF, PCA-SMO, PCA-J48 and PCA-NB are shown in Table IV.

TABLE IV
Detailed accuracy by class: After Feature Extraction

Classification Algorithms	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
PCA-MLP	82%	82%	82%	89%
PCA-RBF	85%	84%	85%	94%
PCA-SMO	82%	80%	82%	94%
PCA-J48	82%	82%	82%	90%
PCA-NB	89%	88%	88%	93%

In this study, the models were evaluated based on the accuracy measures discussed above (classification accuracy, sensitivity and specificity). In our proposed method, we applied PCA-NB algorithm to this dataset. Using this model we achieved a prediction accuracy of 89%. The classification accuracies obtained from the previous studies [11] and this model for hepatitis disease dataset are presented in Table V & VI.

TABLE V
Before feature Extraction: Classification Accuracy and Time

Classification Algorithms	Accuracy	Time
Naive Bayes	84	0.0
Random Forest	83	0.03
Multilayer Perceptron	83	17.94
J48	83	0.03

TABLE VI
After Feature Extraction: Classification Accuracy and Time

Classification Algorithms	Accuracy	Time
PCA-MLP	82	0.48
PCA-RBF	85	0.19
PCA-SMO	82	0.04
PCA-J48	82	0.07
PCA-NB	89	0.01

A. K-Fold Cross-Validation

In order to have a good measure performance of the classifier, k-fold cross-validation method has been used. The classification algorithm is trained and tested k time. In the most elementary form, cross validation consists of dividing the data into k subgroups. 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. We used 10 fold cross-validations in our work.

B. Kappa Statistics

The kappa parameter measures pair wise agreement between two different observers, corrected for an expected chance agreement. For example if the value is 1, then it means that there is a complete agreement between the classifier and real world value. Kappa value can be calculated from following equation

$$K = [P(A)-P(E)]/[1-P(E)] \quad (5)$$

$$P(A) = (TP+TN)/N \quad (6)$$

$$P(E) = [(TP+FN)*(TP+FP)*(TN+FN)]N^2 \quad (7)$$

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. (6). P (E) is the chance of agreement calculated by Eq. (7). In this study the kappa value is 0.6372 for PCA-NB which is calculated by using Eq. (5).

C. Confusion Matrix

A confusion matrix is calculated for Naive Bayes and PCA-NB classifiers to interpret the results. The confusion matrix is shown in table VII and VIII.

TABLE VII
Before Feature Extraction

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

TABLE VIII
After Feature Extraction

a	b	classified as
22	10	a = DIE
8	115	b = LIVE

D. Graph Results

The graph results in Fig 2. Shows performance analysis related to accuracy of various algorithms.

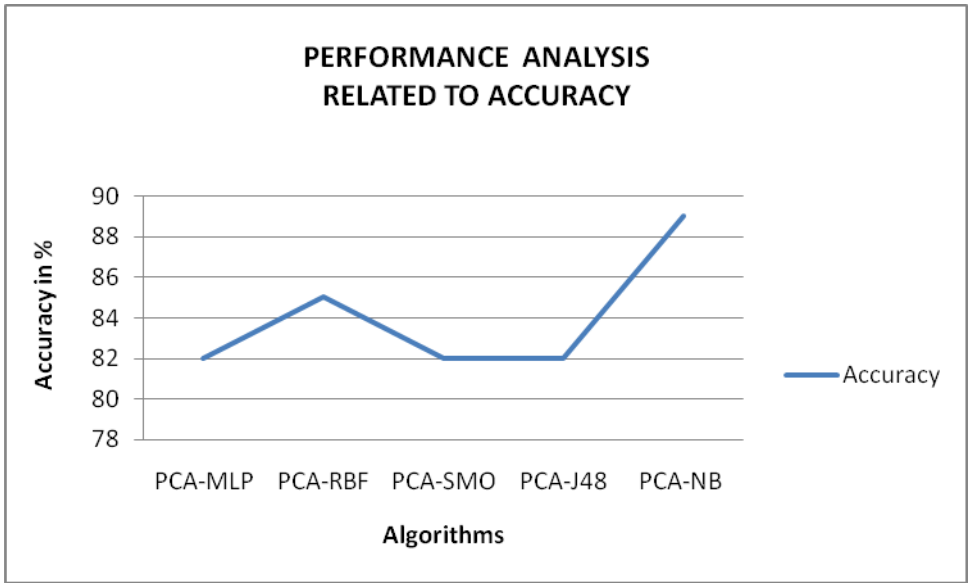

Fig. 2. Performance related to Accuracy

The graph results in Fig 3.shows performance analysis related to time over various algorithms.

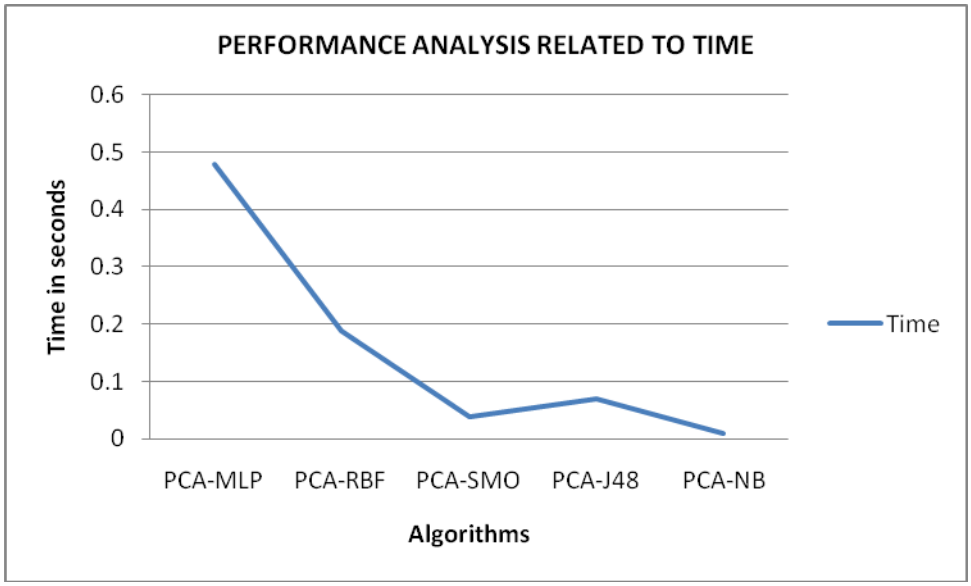

Fig. 3. Performance related to Time

VIII. CONCLUSION

In this work was developed an improved medical diagnostic method called PCA-NB for addressing hepatitis diagnosis problem. Experiments on different portions of the hepatitis dataset demonstrated that the proposed method performed significantly well in distinguishing the live from the dead one. It was observed that PCA-NB achieved the best classification accuracies for a reduced feature subset that contained sixteen features. Meanwhile, comparative study was conducted on the methods of the PCA-MLP, the PCA-SMO, the PCA-J48, the PCA-RBF, PCA-NB and the NB. The experimental results showed that the PCA-NB performed advantageously over the other methods in terms of the classification accuracy and time. We believe the

promising results demonstrated by the PCA-NB can ensure that the physicians make very accurate diagnostic decision. Further the research continues with different medical data set applied for medical diagnosis problem.

REFERENCES

[1] Harun Uguz, "A hybrid approach for text categorization by using x^2 stastics, Principal Component Analysis and partical swam optimization", Academic Journals, vol(8) 37, PP 1818-1828, 2013.

[2] Hyunjin Yoon et al. "Feature Subset Selection and feature ranking for multivariate time series", IEEE Transactions on Knowledge and Data engineering, Vol.17, no. 9, September 2005.

[3] Kemal Polat, Salih Gunes, " Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system", Applied Mathematics and Computation 189(2007) 1282-1291.

[4] Tahseen A. Jilani et.al, "PCA-ANN classification of hepatitis- c patients", International Journal of Computer Applications, volume-14-No.7,Febraury 2011.

[5] Hung lian,"on feature selection with principal component analysis for one-class SVM", Pattern Recognition Letters 33(2012) 1027-1031.

[6] Hlip-Ling Chen Et.al, " A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis" . Expert Systems with Applications 38(2011) 11796-11803. Duygu Calisir and Esin Dogantekin", " A new intelligent hepatitis diagnosis system:PCA-LSSVM", Expert systems with Applications 38(2011) 10705-10708.

[7] Yilmaz Kaya, Murat Uyar, " A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease". Applied Soft Computing 13(2013) .

[8] Huda yasin et al."hepatitis-c classification using data mining techniques", International Journal of Computer Applications, Volume 24, No.3, June 2011.

[9] Javad salami sartakhiti, mohammad Hossein Zangooei, Kourosh Mozafari," Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing(SVM-SA),'' Computer methods and programs in Biomedicine 108 (2012) 570-579.

[10] Diana Dumitru , Prediction of recurrent events in breast cancer using the Naive Bayesian classification, *Annals of University of Craiova*

[11] Math. Comp. Sci. Ser. Volume 36(2), 2009

[12] T. Karthikeyan, P.Thangaraju, "Analysis of classification of algorithm of applied to hepatitis patients", International Journal of Computer Applications, volume-62-No.5, January 2013.