ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

Prediction of Breast Cancer by means of Enhanced Feature Selection and Classification

R.S.Padma Priya [1], P,Senthil Vadivu [2]

Assistant Professor, Department of Computer Technology, Dr.N.G.P.Arts and Science College, Coimbatore, Tamilnadu, India[1]

Head / Associate Professor, Department of Computer Applications, Hindusthan College of Arts and Science , Coimbatore, Tamilnadu, India[2]

ABSTRACT: Medical data processing models are predominant in medical knowledge analysis. This paper introduces the Naïve Bayes (NB) classifier algorithmic rule, together with an improved feature selection technique, towards an accurate finding of breast cancer, in its early stages. Performance of the NB classifier was seen to be noticeably increased once unsuitable features were removed from the modelling method. Empirical verification re-establishes that our hybrid feature-selection approach, using nominal set of attributes, enhances the results obtained from solitary feature selection methods**.**

KEYWORDS: Accuracy, Recursive Elimination, Naive Bayes Classifier, Uni-variate Selection.

I. INTRODUCTION

In recent times, breast cancer is reportable as one of the leading causes of death of middle-aged women. Reports shows that over the new breast cancer cases are found in and round the world. Throughout their lifespan, one in twelve women is predicted to be afflicted by this disease. Like other enervating diseases, cancer is curable if diagnosed and treated in its earlier stages. Unfortunately strikingly massive numbers of the feminine population die because of belated detection of this disease. This state of affairs has alternated noticeably in last few years and to the current issue numerous resolutions are explored, however with an oversized set of features. Therefore, a data-mining model was planned to warn borderline cancer patients with a more accurate prediction, supplying speedy and cost effective solution. For conceiving this model, a systematic investigation was performed for a comparative study in up the prediction accuracy. Within the first part, Naive Bayes algorithmic rule was applied to data set; within the second part, feature selection methods are applied so as to eliminate the options, which were clumsy for the prediction. once feature selection method(s) were used, an improvement was determined in the performance of the NB classifier.

Apart from the already existing feature selection methods, we have introduced a new hybrid technique, which is a combination of feature selection method(s) like uni-variate selection and recursive elimination, so as to extract the most distinguished features. In this data-mining model, the prominent features obtained from hybrid approach can be used to study the diagnostic details of cancer for a good treatment.

II. RELATED WORK

In [1] the author's objective was to work out whether continuous feature selection technique for the sliding window technique of stream mining, based on Naive Bayes, leads to improved performance. In Stream mining, the feature selection technique is finished using same computational information structures and it leads to fast and efficient implementation of Naive Bayes classification.

In [2] the authors proposed a model to enhance the Naive Bayes classification performance employing a new Artificial Immune System (AIS) supported self-adaptive attribute weighting. It has been determined that attributes utilized in learning task are related to every alternative because of which Naive Bayes conditional independence assumption

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

decreases its classification performance. AIS methodology assigns correct weight values for NB classification. With the assistance of AIS weight values best attributes are learned for NB classification.

In this paper, we tend to use a combined approach in feature selection method(s) to get distinguished features so as to study and observe the modification in performance of NB classifier. This type of combination approach in feature selection strategies applied in data mining model for earlier detection of cancer haven't been studied or approached to our knowledge.

III. SYSTEM DESIGN

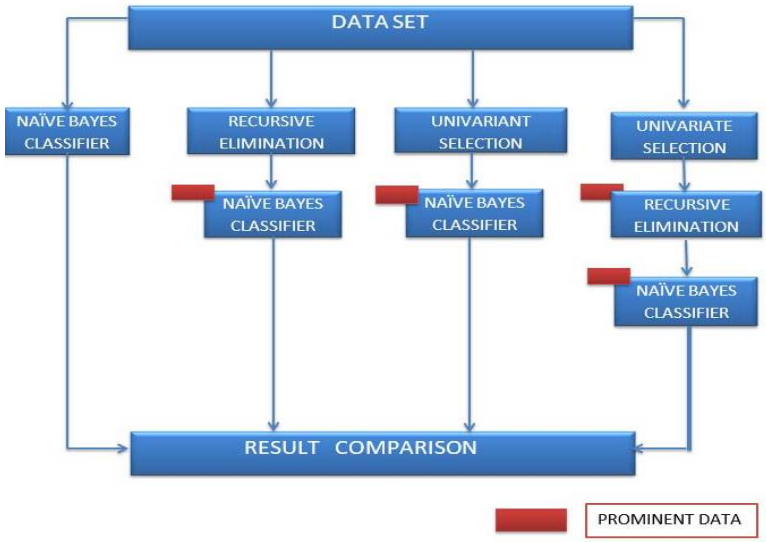

Figure 1. Overall System Flow

We are using breast cancer/carcinoma data set (BCDS) to be handled by our proposed data-mining model, which will classify cancerous /non-cancerous details using Naive Bayes algorithm. In data mining technique we processes raw data into useful data and applied algorithm(s) to search out pattern, generate predictions and create inferences regarding the data. Our overall system style is shown in Figure 1 .

3.1. Phase 0

Our BreastCancerDataSet raw data set comprises 11 attributes where attributes defines the property of an object and its deciding factor for the possibilities of prediction of cancer. Here the selected attributes were family tree, breast feeding, OCP, Auxiliary lymph node status, ultrasound (USG), Mammogram, True Cut biopsy, Biopsy, HER2 status, ER, Diagnosis. From these attributes/features, we discover distinguished features using feature selection method(s), which is able to provide the cancer details and symptoms.

3.2. PHASE 1

Using sklearn.naivebayes.GaussianNB module from Scikit-learn Naive Bayes algorithm has been applied on BreastCancerDataSet. Scikit-learn is a software consisting of library for the Python programming language.

3.3. Naive Bayes Classifier:

Naive Bayes algorithmic program provides the probability of an object with bound variety of features belonging to a specified class. it's a probabilistic classifier based on Bayes theorem with the idea of independent features. NB algorithm has a straightforward and easier approach on real time data like BCDS. The mathematical equation followed in NB algorithmic rule is represented below:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

$$(Ci \mid d) = (Ci)p(d \mid Ci)p(d) \quad (1)$$

Where $(Ci|d)$ is the posterior probability of class Ci given a brand new dependent feature vector d, p(Ci) is that the probability of class Ci which can be calculated by

$$(Ci) = Ni/N \quad (2)$$

Where Ni is the range of dependent feature vectors allotted to class Ci and N is the range of classes, p(d/Ci) is the probability of a dependent feature vector d given a class Ci and p(d) is the probability of dependent feature vector d.

Naive Bayes algorithmic rule are handling BCDS and distinguished BCDS (BCDS with relevant features) in 3 completely different phases of our experiment .The data set (BCDS) is organized into attributes and class label when using in classifier. For training and testing purpose the information set is split into 70% and 30% respectively. Proposed work on this model involves scheming, observing and learning on following aspects once NB algorithms is applied with completely different feature selection strategies

3.4. CLASSIFIER PERFORMANCE MEASURES:

We used the idea of TP, TN, FP and FN for getting accuracy, preciseness and recall details. They are defined as follows:

1. **True Positive(TP):** a real positive check result is one that detects the condition when the condition is present .

2. **True Negative(TN):** a real negative check result is one that doesn't discover the condition once the condition is absent.

3. **False Positive(FP):** A false positive check result is one that detects the condition when the condition is absent.

4. **False Negative(FN):** A false negative check result is one that doesn't discover the condition when the condition is present

5. **Accuracy :**It is the most intuitive performance measure and it's a ratio of correctly foretold observation to the total observations.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

6. Sensitivity (Recall): measures the flexibility of a test to detect the condition once the condition is present.

$$Recall = TP\ (TP + FN) \quad (4)$$

7. Exactness: Precision is the ratio of exactly predicted positive observations to the total predicted positive observations.

.

$$Precision = TP\ (TP + FP) \quad (5)$$

3.5. PHASE 2

We use two feature selection method(s) in order to obtain distinguished features from BCDS. Feature selection is a method of extracting relevant features from the data set for use in model construction. The initial data set, BCDS is fed to those Feature selection method(s) to process and obtain distinguished features. After this, distinguished data set(s) which include chosen features and class label (Attribute: Diagnosis) is fed to Naive Bayes classifier and also the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

observations are recorded. We have in brief represented the two feature selection method(s) that is used to acquire distinguished data set.

The feature selection method(s) are:

3.5.1. UNI-VARIATE SELECTION:

:In this methodology we have a tendency to use the Select K-Best method in which we assign a value for k, that means k range of features are extracted from the given data set. To implement this we will score all the features using a mathematical relation referred to as the Chi-Squared, which is a statistical test to find necessary features from a data set.

Here the value given for k is 4, that is it will extract 4 out of the best features from BCDS . after which a distinguished data set of BCDS is created with class label along with these 4 selected features and fed to NB classifier. The Chi-Squared mathematical equation is given as:

$$X2(f, t) = N(AD - CB)2\ (A + C)((B + D)(A + B)(C + D)\ (6)$$

f is a feature, t is a target variable that we want to predict, A is the range of times that f and t co-occur, B is the number of times that f happens while not t, C is that the range of times that t occurs without f, D is the number of times neither t or f occur and N is the number of observations.

3.5.2. RECURSIVE ELIMINATION:

Our main aim is to get the most distinguished features from the data set. in this methodology it removes the lowest ranking features recursively and this is often done using logistic regression. Logistic regression is applied on feature set and a coefficient value for every feature is found. The feature with least coefficient value is graded lowest which will be removed and again regression is applied on remaining of the features. it is done until all the features are exhausted and features are graded consequently. From this we obtain distinguished features and we feed a new table, which contains obtained distinguished features along with class-label to Naive Bayes classifier again, to note the value for additional comparative study. This technique uses the logistic Regression and the formula is given by:

$$lo(odds) = p(x)\ 1 - p(x)\ (7)$$

3.6. PHASE 3

In this phase of our experimental work we have performed an hybrid approach of using distinguished feature(s) obtained by both Uni-variate selection method and recursive Elimination method. The original data set is fed first to Uni-variate selection from which a set of distinguished features will be obtained. These obtained distinguished features are then handled by recursive Elimination method to refine minimum features that is having a lot of relevancy. After this, distinguished data set (prominent features along with class labels) is created and fed to Naive Bayes algorithm and the observation is noted for a comparative study as shown in Figure 2. PF1 and PF2 denote distinguished features after the first and second step respectively.

.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

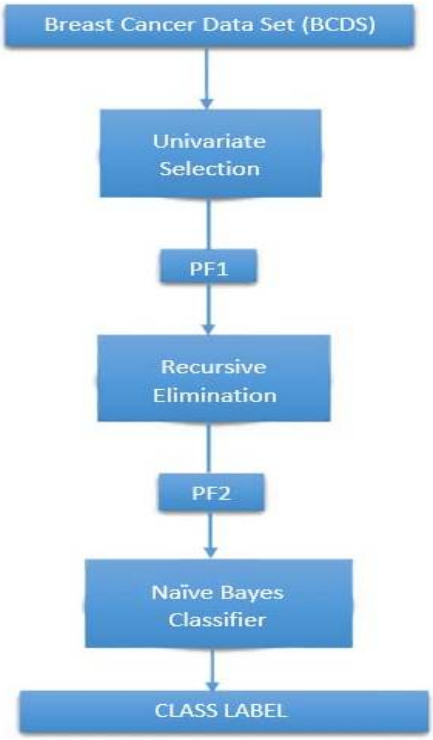

FIGURE 2:IMPROVED APPROACH

IV.EXPERIMENTAL RESULTS

Module training	Accuracy	Execution time
Naive Bayes	0.8293411	10.106014
Uni-variate selection	0.8383233	0.6906962
Recursive Elimination	0.8413173	0.6771259
Hybrid methodology	0.8483233	0.6798429

Table 1. Accuracy Values and Execution Times

In Table 1 we have recorded the training accuracy and execution time in each phase i.e Naive Bayes, Uni-variate Selection, recursive selection and Hybrid.

Module training	Precision	Recall
Naive Bayes	0.79	0.82
Uni-variate selection	0.80	0.84

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

Recursive Elimination	0.80	0.85
Hybrid methodology	0.81	0.81

Table 2. Classification Performance Measures: Training Set

Module training	**Precision**	**Recall**
Naive Bayes	0.82	0.82
Uni-variate selection	0.81	0.78
Recursive Elimination	0.81	0.79
Hybrid methodology	0.84	0.79

Table 3. Classification Performance Measures : Testing Set

Module	**Feature(s)**
Uni-variate selection	Family tree, Auxiliary lymph node status, Mammogram, HER2 status
Recursive Elimination	Family tree, Auxiliary lymph node status, USG
Hybrid method	Family tree, Auxiliary lymph node status, HER2 status

TABLE 4. Distinguished features

The execution time acquire in Naive Bayes is 10.10 ms and when applied Feature selection method(s) we identified a considerable modification in execution time to an average of 0.68ms from 10.10 ms. Similarly, training accuracy enhanced from 0.829341 of Naive Bayes to 0.84131736 of recursive Elimination. When hybrid technique was applied accuracy increased to 0.83 and execution time to 0.67984 ms. Therefore when used relevant features which is obtained from feature selection method(s), NB classifier enhanced its performance both in accuracy score and execution time.

In Table 2 the precision value changed from 0.79 of Naive Bayes to 0.81 of Hybrid methodology. Similarly the recall value changed from 0.82 of Naive Bayes to 0.81 Hybrid. The True Positive (TP) score for Hybrid is 118 and Naïve Bayes 119 just in case of training Set. The F-measure value for Naive Bayes and Hybrid are 0.805 and 0.81 respectively.

In Table 3 the precision value changed from 0.82 of Naive Bayes to 0.84 Hybrid methodology. The Recall value changed from 0.82 of Naive Bayes to 0.79 Hybrid. The True Positive (TP) score and False Negative (FN) is 61 and 16 respectively for Hybrid and True Positive (TP) score and False Negative (FN) is 63 and 14 respectively for Naïve Bayes. The F-measure value for Naive Bayes and Hybrid are 0.82 and 0.817 respectively.

Accuracy provides the classification score for classification system thereby, precision and recall performance measure are used for verifying predictability. Accuracy obtained for Hybrid is 0.8483233 and Recall obtained for Hybrid is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 6, June 2019

0.81. it's observed once used hybrid approach for the classification system the accuracy score and the recall value are close to one another.

In Table 4 we obtained minimum features/attributes in each feature selection method(s). These are the distinguished features out of the initial features in BCDS. These features can be used to study and find valid symptoms behind the cause of totally different types of cancer, thereby helping the patients and doctors to proceed with treatments in initial stage of cancer.

V. CONCLUSION

In the proposed work a new hybrid feature selection methodology for earlier prediction of breast cancer has been identified. Our data mining model uses a hybrid feature selection method that is highly efficient, accurate and surpasses the results obtained from individual feature selection strategies. This helps to diagnose cancer in the initial stage and can have positive effect on the treatment and additionally in curability.

REFERENCES

[1] Patricia E.N. Lutu , Fast Feature Selection for Naive Bayes Classification in Data Stream Mining , Proceedings of the World Congress on Engineering 2013 Vol III.
[2] Jia Wu ,Shirui Pan ,Xinguan Zhu ,Zhihua Cai ,Peng Zang ,Chengqi Zhang , Self-adaptive Attribute Weighting For Naive Bayes Classification , Expert Systems With Applications 2(2015) 1487-1502).
[3] C. P. Prathibhamol, Ashok, Solving multi label problems with clustering and nearest neighbor by consideration of labels, Advances in Intelligent Systems and Computing, vol. 425, pp. 511-520, 2016.
[4] Adeena K D, Remya R , “Extraction of relevant dataset for support vector machine training: A Comparison”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, pp. 222-227, 2015.
[5] Krithika, R. and Narayanan, Jayasree , “Learning to Grade Short Answers Using Machine Learning Techniques”,Proceedings of the Third International Symposium on Women in Computing and Informatics , pp.262–271, 2015.
[6] J. Bhaskar, Sruthi, K., and Prof. Nedungadi, P. Hybrid approach for emotion classification of audio conversation based on text and speech mining , in Procedia Computer Science, 2015.
[7] Krishnaveni K S ,Rohit R Pai, Vignesh Iyer ,Faculty Rating System Based on Student Feedbacks Using Sentimental Analysis,2017 International Conference On Advances in Computing,Communications and informatics,ICACCI 2017.
[8] M. Raniszewski, "Sequential reduction algorithm for nearest neighbor rule", Computer Vision and Graphics, 2010.
[9] A. Osareh and B. Shadgar, “Machine learning techniques to diagnose breast cancer,” in Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics (HIBIT’10), pp. 114–120, April 2010.
[10] M. Perez and T. Marwala, “Microarray data feature selection using hybrid genetic algorithm simulated annealing,” in Proceedings of the IEEE 27th Convention of Electrical and Electronics Engineers in Israel (IEEEI ’12), pp. 1–5,November 2012
[11] N. Revathy and R. Balasubramanian, “GA-SVM wrapper approach for gene ranking and classification using expressions of very few genes,” Journal of Theoretical and Applied Information Technology, vol. 40, no. 2, pp. 113–119, 2012
[12] M.H. Asri, H.A Moatassime, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". Procedia Comput Sci, Vol. 83, pp. 1064–1073, 2016.
[13] A. Alarabeyyat, A.M., "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm", in 9th International Conference on. IEEE, v.i.e.E. (DeSE), pp. 35-39, 2016.
[14] M. Morovvat and A. Osareh, “An Ensemble of Filters and Wrappers for Microarray Data Classification” Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.2, pp. June 2016.
[15] Q. Su, “A Cancer Gene Selection Algorithm Based on the K-S Test and CFS”,Biomed Research International, pp. 1-6, 2017.
[16] S.K. Prabhakar, H. Rajaguru, "Performance Analysis of Breast Cancer Classification with Softmax Discriminant Classifier and Linear Discriminant Analysis", In: Maglaveras N., Chouvarda I., de Carvalho P. , Springer 2018.