NET-EFFECT IN THE DIAGNOSIS OF BREAST CANCER USING FEATURE REDUCTION AND CLASSIFICATION

K. Vijaya Sri[1], K. Usha Rani[2]

[1]*Research Scholar: Department of Computer Science, SPMVV (Women's University), Tirupati, India*
[2]*Research Supervisor: Department of Computer Science, SPMVV (Women's University), Tirupati, India*

Abstract

Soft Computing methods provide solutions to biologically inspired problem of medical domain like breast cancer. Neural Networks, Fuzzy Logic and Genetic Algorithms contribute novel algorithms to deal with breast cancer. Breast cancer can be diagnosed using soft computing methods. In this paper, we try to produce effective diagnosis of breast cancer by using feature reduction and classification methods. The net-effect of the classification before and after feature reduction process is stated. The feature reduction method applied is Principal Component Analysis (PCA) and the classification method includes Support Vector Machines (SVM). The result of the proposed method produced better outcome when applied on Wisconsin Breast Cancer Data Set (WBCD).

Keywords— *Soft Computing, Neural Networks, Breast Cancer, feature reduction, PCA, Classification, SVM.*

----------***----------

1. INTRODUCTION

Uncertain and precise real world problems having natural existence are solved using Soft Computing Techniques. Combined use of soft computing techniques results in better solutions to the problems [1]. Breast Cancer is the second leading cause of death in women worldwide. One-third of the global breast cancer burden accounts on US, India and China. The study on breast cancer now-a-days is observed to be in male and female.

World Health Organization provides the statistics in 2012 that for the years 2015, there will be an estimated 1, 55,000 new cases of breast cancer and about 76,000 women in India are expected to die of the disease [2].

American Cancer Society 2015-16 facts and figures states that in US 292,130 women suffer from breast cancer and additionally 40,290 women face death due to breast cancer [3].

In China as per the 2011 statistics, out of 75,310,169 populations, 248,620 breast cancer cases were identified [4]. It is suggested that breast feeding and physical activity are the two factors that make women away from the risk of breast cancer. The incidence of breast cancer can be easily identified through diagnosis of the clinical samples of the patients.

Breast cancer based study was earlier carried mostly based on data mining [5], [6]. Recently it includes many contributions of soft computing. Neural networks are widely used to work with breast cancer. It is observed that some applications give better results based on hybrid methods [7]. The organization of the study is Section-II contains Literature Review, Section-III contains Wisconsin Breast Cancer Dataset, Section-IV contains Brief note on the theory of methods applied, Section-V Proposed Methodology, Section-VI contains Performance Measures, Section-VII contains Experimental Results and Section-VIII contains Conclusions of the study.

2. LITERATURE REVIEW

Literature states that the efficiency of the classification in breast cancer can also be improved by feature extraction [8], [9]. Literature on breast cancer using PCA and SVM highlight various dimensions and nature of work performed with respect to area of research and on different datasets. It is observed that most of the PCA and SVM works are carried on Image Processing. Related to clinical samples research shows that by applying PCA and SVM on WBDC dataset 0.9681% of test accuracy is obtained [10]. PCA and Fuzzy-SVM used to speed up training and testing of the classifier and testing correct rates are taken for expert doctors for re-examination [11]. Based on these literatures a method is in the use of PCA and SVM using WBCD.

3. WISCONSIN BREAST CANCER DATASET

UCI Machine Learning Repository provides datasets to perform the empirical analysis of machine learning [12], [13]. Dataset contains 699 instances and 10 attributes defining the features of breast cancer. The features include:

No.	Attributes	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 – 10
3.	Uniformity of Cell Size	1 – 10
4.	Uniformity of Cell Shape	1 – 10
5.	Marginal Adhesion	1 – 10
6.	Single Epithelial Cell Size	1 – 10
7.	Bare Nuclei	1 – 10

8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	(benign-2, malignant-4)

There are 16 instances of missing attribute value. The missing values are replaced by the mean of the attribute values. The first attribute is removed as it has no significance in its usage. Hence the dataset of the trial includes 9 features distributed within the interval of 1-10 whereas the 10^{th} attribute is the class which can be categorized as benign that holds a value of 2 or malignant holds the value 4. This data set contains 65.5% benign data and 34.5% of malignant data.

4. BRIEF NOTE ON THE THEORY OF METHODS APPLIED

The methods used in this study include feature reduction by Principal Component Analysis and classifier used is Support Vector Machine.

4.1 Principal Component Analysis

Principal Component Analysis plays a vital role in order to reduce x-dimensional data into (p)-dimensional data where (p<x) to produce efficient results with less computational efforts. The principal components are the eigenvectors having the eigenvalues as their magnitude, obtained from covariance matrix. Eigenvalues are sorted in decreasing order and the data points that are having larger magnitude are identified. Feature reduction is done when the eigenvalues have larger magnitude than others. Values that preferred least are the larger dimensional points and are neglected as data features. Choose (p)-dimensions (where p<x) having prominent sub-space. Finally projection matrix of (p)-dimension subspace is considered. Then the original dataset is transformed into p-dimensional projection matrix.

4.2 Support Vector Machine

Support Vector Machine is a statistical method of classification using training and testing samples. Support Vector Machine with Radial Basis Function (RBF) kernel is used which is popular due to its localized nature and finite response behavior. Support Vector specifies a decision boundary of optimal plane. Support vector tries to maximize the separation of a particular hyper plane. Finally it identifies the optimal plane. Kernel of a support vector specifies the mapping of the points in the plane. RBF kernel function of Gaussian type has the following calculation:

$$K (X_i , X_j) = \exp (-\gamma | X_i - X_j |^2)$$

where X_i is the input dimension, X_j is the output dimension, γ is the adjustable parameter.

Various tools support the experimentation of soft computing methods [14]. In this study the experiments are conducted using the tool MATLAB R2013a [15].

5. PROPOSED METHODOLOGY

The proposed method uses Support Vector Machine and Principal Component Analysis. SVM gives statistically better results when trained and used. PCA gave better results in most of the applications related to medical domain [8],[9]. These two methods are chosen to specify their effect when applied to WBCD cancer dataset. The proposed algorithm *PCA_SVM* classifies breast cancer data of WBCD in a staged manner. Stages of solution are depicted in the proposed algorithm.

Algorithm: PCA_SVM
STAGE-1
Step-1: Load WBCD dataset of 699 instances
Step-2: Pre-Process the dataset
Step-3: Partition the database
Step-4: Train the network using SVM parameters.
Step-5: Save output_1
Step-6: Load the 9-dimensional data file
Step-7: Feature reduction by PCA
Step-8: Save low_dim
Step-9: Train using SVM on low_dim
Step-10: Save output_2
STAGE-2
Step-1: Analyse produced results.
Step-2: Net-effect and significance specification

The proposed algorithm is estimated to give better results for breast cancer diagnosis.

6. PERFORMANCE MEASURES

The performance of the experiment is calculated based on the measures of Accuracy, Sensitivity and Specificity. It also includes the calculations of False Positive (FP) rate, True Positive (TP) rate and Negative Matrix (NM) rate. The performance analysis is derived based on the confusion matrix as shown in Table 1. The predictors in the columns are compared with the actual class of the rows or vice-versa.

Table 1: Confusion Matrix

Total Population		**True Condition**	
		Positive	**Negative**
Predicted Condition	**Positive**	True Positive(TP)	False Positive (FP)
	Negative	False Negative(FN)	True Negative(TN)

Accuracy [AC]: Total no. of correct predictions of the samples.
True Positive [TP] Rate: Correctly identified positive instances.
False Positive [FP] Rate: Incorrectly identified positive instances.
True Negative [TN] Rate: Correctly identified negative instances.
False Negative [FN] Rate: Incorrectly identified negative nstances.

The following calculations are considered to specify the result of classification:

Classification Accuracy, AC = (TP+FP) / (TP+FP+FN+TN)

Sensitivity = TP / (TP+FN) %

Specificity = TN / (FP+TN) %

7. EXPERIMENTAL RESULTS

Data partitioning is a method in which data is categorized into two sets separately during the process of learning. The train set of WBCD is chosen as 80% and test set as 20% of the data. SVM is experimented on train set and test set. Total instances considered are 699.

7.1 Stage-1

Proposed algorithm for STAGE-1 is experimented and the results are compared with SVM alone. The figures of the experiment are tabulated and are shown in Table II.

Table 2: Experimental Results

STAGE-1	***SVM***	***PCA_SVM***
Accuracy	0.65	0.65
Sensitivity%	0.01	0.01
Specificity%	0.0857	0.092
Error Rate	0.0576	0.0288

The plot of correctly classified and incorrectly classified instances is specified by their support vectors. The X-axis contains the input samples and the Y-axis contains the response factors. Data points constituting the support vectors are encircled and the optimal plane is shown in Fig. 1.

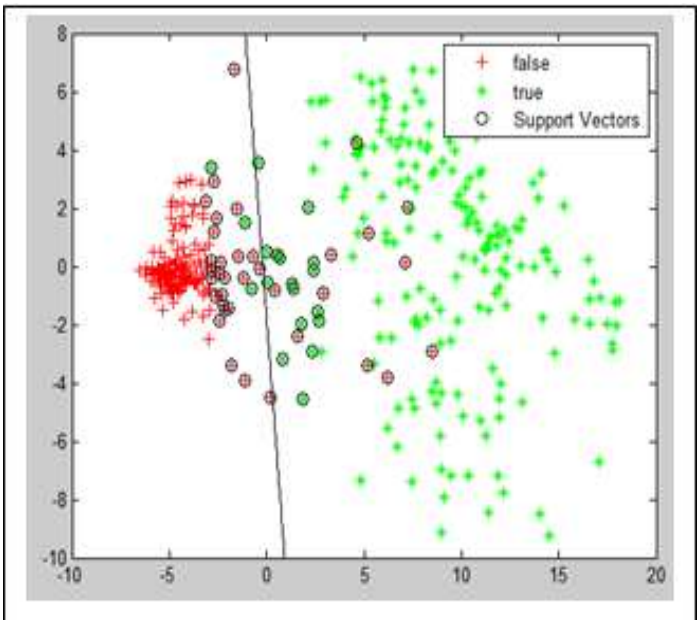

Fig 1 Support Vectors with feature reduction

7.2 Stage-2

The analysis of support vectors generated and the tabulated analysis parameters are evaluated. SVM has 94.24% classification accuracy whereas by executing PCA and SVM the classification has 97.12% accuracy. The net-effect of the proposed algorithm states that error rate can be reduced by 0.0288.

8. CONCLUSION

Breast cancer needs to provide accuracy of diagnosis at the early stages. Even a variation in negligible value of diagnosis may result in better contributions to the experts when used in support with the technology. In this study, the classifier SVM alone and along with PCA is experimented on breast cancer dataset. It is concluded that the proposed algorithm shows betterment in the diagnosis. The work may also be extended by comparing all other feature reduction methods available.

ACKNOWLEDGMENT

The Breast Cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

REFERENCES

[1] Amit Konar, *Artificial Intelligence and Soft Computing* Behavioral and Cognitive Modeling of Human Brain. CRC Press LLC, 2000.

[2] http://www.breastcancerindia.net/statistics/stat_global.html.

[3] American Cancer Society. Breast Cancer Facts & Figures 2015-2016. Atlanta: American Cancer Society, Inc. 2015.

[4] Jie He. National Office for Cancer Prevention and Control, National Cancer Center, Beijing 100021, Chin J Cancer Res. 2015 Feb; 27(1): 2–12.Chinese Journal of Cancer Research ISSN 1000-9604.

[5] D.Lavanya , Dr.K.Usha Rani A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 1, January 2013 ISSN 2319 – 4847.

[6] D.Lavanya, Dr.K.Usha Rani "Ensemble Decision Tree Classifier for Breast Cancer Data"International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.

[7] K. Vijaya Sri, Dr. K. Usha Rani "Neuro-Fuzzy Systems and Applications – A Review" International Journal Publications of Problems and Applications in Engineering Research, CSEA2012, Vol 04, Special Issue 01; Pg. No's 202-205, 2013 , ISSN: 2230-8547; e-ISSN: 2230-8555.

[8] G. Naga Rama Devi, Dr. K. Usha Rani "Importance of Feature Extraction for classification of Breast Cancer Datasets – S Study", International Journal of Scientific and Innovative Mathematics Research(IJSIMR), Vol. 3, Special Issue 2, July-2015, PP 763-768, ISSN 2347-307X.

[9] G. Naga Rama Devi, Dr. K. Usha Rani "Evaluation of Classifier Performance using Resampling on Breast Cancer", International Journal of Science and Engineering Research, Vol. 6, Issue 2, February 2015, ISSN 2229-5518.

[10] Xiu-feng Yang, Hui Peng Xiao-feng Zhou, Young-Iai. Zhang, Sheyang Institute of Automation, Chinese Academy of Science, Beijing, China.

[11] Zhaohui Luo, Xiaoming Wu; Shengwen Guo; Binggang Ye, Diagnosis of Breast Cancer Tumor Based on PCA and Fuzzy Support Vector Machine Classifier, Coll. Of Biol. Sci. & Eng., South China Univ. of Technol., Guangzhou; Natural Computation, 2008. ICNC '08. Fourth International Conference on (Volume:4).

[12] UCI Machine learning Repository, http://archive.ics.uci.edu/ml/.

[13] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[14] K. Vijaya Sri, Dr. K. Usha Rani "Ingenious Tools of Soft Computing", International Journal of Engineering Sciences Research-IJESR http://ijesr.in/ ACICE-2013 Vol 04, Special Issue 01, Pg. No. 1305-1312, 2013 ISSN: 2230-8504; e-ISSN-2230-8512

[15] http://in.mathworks.com/