International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

A Model for Classification of Wisconsin Breast Cancer Datasets using Principal Component Analysis and Back-Propagation Neural Network

Shweta Saxena [1], Manasi Gyanchandani [2]

[1, 2]Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

Abstract: ***Nowadays, the second leading reason of death (due to cancer) among females is breast cancer. Early detection of this disease can significantly enhance the probabilities of long-term survival of breast cancer patients. This paper proposes a computer-aided-diagnosis model for Wisconsin Breast Cancer (WBC) datasets using Back-Propagation Neural Network (BPNN). The data pre-processing technique named principal component analysis (PCA) is proposed as a feature reduction and transformation method to improve the accuracy of BPNN.***

Keywords: Breast Cancer, Computer-Aided-Diagnosis, Principal Component Analysis, Back-Propagation Neural Network

1. Introduction

Breast cancer is the common cancer in women after the lung cancer and a major cause of death [1]. Cancer is a disease in which cells develop abnormally. In breast cancer, an uncontrolled cell growth initiates in tissues of the breast and form tumor. It may invade neighboring tissues and may extent to the lymph nodes and further parts of the body [3]. The breast cancer cases are 20 percent of the total cancer cases in world. Females in the world face the exclusive challenges to protect themselves from this disease. They have a lesser screening rates as compared to the general population. A new research found a high occurrence of breast cancer in medicated female patients and a significantly greater health care use and expenses [2]. There is much research on principal component analysis (PCA) data preprocessing and diagnosis of breast cancer with neural network on WBC data. In [4] a pre-processing model using component analysis is proposed. The authors used Kernel K-mean algorithm and Expectation Maximization to cluster the pre-processed data. The proposed model was tested with help of the lung cancer dataset, Wisconsin Breast Cancer (WBC) dataset, and prostate cancer dataset. According to the results, the performance of the cluster vector value is high with a short processing time. In [11] the problem of PCA learning in the existence of missing values is reviewed. According to the authors, the PCA with probabilistic formulation gives a good basis for handling the missing values. Over fitting becomes a severe problem for the high dimensional and very sparse dataset, and traditional PCA algorithm is very slow. The authors introduced a Bayesian learning based fast algorithm. In [14], the performance of the statistical NN structures, radial basis network, general regression neural network (GRNN) and probabilistic NN were examined on the WBC dataset. According to the results, GRNN gives the best classification accuracy when the test set is considered. In [15] it was found that the recent use of the combination of Artificial NN most of the instances gives accurate results for the diagnosis of breast cancer. In [16], a feed forward neural network is built and the back-propagation (BP) algorithm was used to train the network. Among the six methods, Levenberg Marquardt method gave the highest result with 99.28% accuracy. A min-max normalization based pre-processing was used in this research. The study [20] presents a comparative analysis of different ensemble learning methods for classifying the Wisconsin breast cancer dataset. In [21] the authors used Breast Cancer Wisconsin (Prognostic) dataset for training and testing a holo entropy enable decision tree. In [22] the researchers addressed the limitation of Select and Test (ST) based medical reasoning algorithm. The researchers first created an efficient input mechanism to read, filter and clean the input from the Wisconsin data. Then, semantic web languages (ontologies and rule languages) were used to make a coordinated rule set. After that, a knowledge representation framework was made to aid the reasoning algorithm. In [23], the subset method was proposed for improving the single-output Chebyshev-polynomial neural network (SOCPNN). The researchers achieved a 100.00% testing accuracy in classification of the WBC data.

In this paper we proposed a Breast Cancer diagnosis model using PCA & BPNN which can be applied on two different WBC datasets. This paper is organized as follows: Section 2 presents the dataset description. Section 4 presents the proposed model, and section 4 concludes the paper.

2. Dataset Description

The Wisconsin Breast Cancer datasets are publicly available in the Internet provided by university of Wisconsin hospital, Madison from Dr. William H. Wolberg. Two different datasets named WBC (original) dataset and Wisconsin Diagnosis Breast Cancer (WDBC) dataset can be used for breast cancer diagnosis.

2.1 Wisconsin Breast Cancer Dataset (Original) [6][7]

This dataset contain 699 samples along with 10 attributes

(including the class attribute). Attribute 1 through 9 represent one complete sample. Each sample has one of the two possible classes: benign or malignant. Total 458 (65.5% of total instances) are benign and 241 or 34.5% samples are malignant. The attribute information is provided by Table 1.

Table 1: Wisconsin Breast Cancer Dataset (Original)

S.no	*Attribute*	*Range*
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10
6	Bare nuclei	1-10
7	Bland chromatin	1-10
8	Normal nucleoli	1-10
9	Mitosis	1-10
	Class	2 (benign) or 4 (malignant)

2.2 Wisconsin Diagnosis Breast Cancer (WDBC) [6][7]

This database has 569 samples and 31 attributes including the class attribute. Attribute 1 through 30 represent the complete samples. Each instance belongs to one of the two possible classes: benign or malignant. According to the class distribution, 357 samples are Benign whereas 212 instances are Malignant. The details of the attributes of WDBC database is: ID number, Diagnosis (M = malignant, B = benign) and 10 real-valued feature descriptors which are computed for the nucleus of each cell: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension [8] [9]. These feature descriptors are calculated from a digital image of a fine needle aspirate (FNA) sample of a breast mass.

3. Proposed Breast Cancer Diagnosis Model

Pre-processing transforms the data into simple and an effective form. PCA is unsupervised learning method for data preprocessing [4].

In PCA, an orthogonal transformation is used to convert a given set of observations which contain possibly correlated variables into a linearly uncorrelated variable set. These variables are called the principal components. The number of principal components should be less than or equal to the number of the source variables [10]. PCA finds linear transformations of data retaining the maximal amount of variance [11]. The PCA can be used as a features reduction and transformation process which combines a set of correlated feature descriptors [10]. Fig 1 shows data preprocessing using PCA. PCA reduces the dimensions of the data while retaining as much as possible of the variation present in the original dataset. The best low-dimensional space can be determined by the best eigenvectors of the covariance matrix of M. The eigenvectors corresponding to the highest Eigen values are also called principal components.

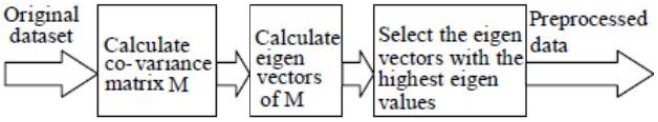

Figure 1: Data pre-processing using PCA

The principal eigenvectors which are orthogonal represent the directions the maximum value of the signal. This property improves the system performance by speeding up the convergence of model training. The feature space has the reduced set of features that actually contributes to classification. It cuts the pre-processing costs and minimizes the 'peaking phenomenon' effects in classification [17]. After pre-processing the WBC data can be applied to BPNN which classifies the data into two sets- benign and malignant. Fig 2 shows proposed WBC data preprocessing and diagnosis model. The overall process consists of 2 phases- WBC data preprocessing and applying the preprocessed data to NN for classification. PCA involves feature extraction and selection. Feature selection finds a subset of the source variables and reduces and eliminates the noisy dimension. Noisy or irrelevant features can have the same effect on classification as other predictive features therefore they creates a negative impact on accuracy [18]. Feature extraction transforms a high-dimensional dataset into lower dimensional dataset [4]. Advantages of feature reduction [19] includes the identification of a reduced set of features that are predictive of outcomes can be very useful from a knowledge discovery perspective. For many learning algorithms, the training and/or classification time increases directly with the number of features, which is efficiently reduced by dimension reduction methods. Advantages of NN include a great tolerance to noises data and the ability to classify the unseen patterns. Other advantages of NN include adaptive learning, self-organization, real time operation, and fault tolerance [5].

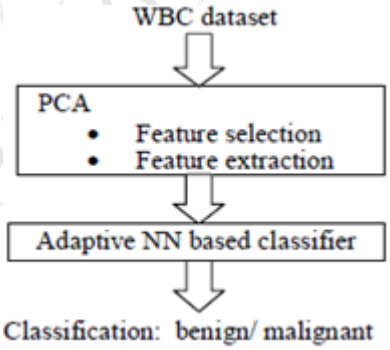

Figure 2: Proposed breast cancer diagnosis model

Back Propagation (BP) is one of the most important discovery in neural networks. The networks that use BP learning algorithm are called BPNN [5]. A BPNN is a multilayered, feed-forward network comprised of an input layer, one or more than one hidden layers, and the output layer. The neurons in the hidden and output layer have biases. These biases are the connections whose activation is always 1. The bias also acts as weights for a training set containing input-output pair. A BP learning algorithm is a procedure for updating the weights in a BPNN for correctly classifying the input patterns. This is a method where error is

propagated back to the hidden unit. The error is the difference between the actual (calculated) and desired (target) output [10]. According to the BP algorithm, the input and output of the neuron, i, (except for the input layer) [13] can be formulated as equation (1) and (2).

$$\text{Input} \quad X_i = \sum W_{ij} O_j + b_i \qquad (1)$$

$$\text{Output} \quad O_i = f(X_i) \qquad (2)$$

Where W_{ij} is the weight of the link from neuron i to node j, b_i is the bias, and f represents the activation function. The summation in (1) is over all neurons, j, in the preceding layer. The output function is a nonlinear function [12].The training algorithm and various parameters used for training BPNN is as follows [5]

Input training vector x = $(x_1, \ldots, x_i \ldots, x_n)$

Output target vector t =$(t_1, \ldots t_k, \ldots t_m)$

δ_k= error at the output unit y_k

δ_j= error at the hidden unit z_j

α= learning rate

V_{oj}= bias on hidden unit j

z_j= hidden unit at j

w_{oj}= bias on output unit k

y_k= output unit k.

3.1 Training Algorithm

1) Assign initial values of weight (small random values).
2) Do steps 3-10 while stopping condition is false.
3) Do steps 4-9 for each training pair.
4) All input unit accepts the input value x_i and transmit to all the hidden units.
5) Each hidden unit (z_j,j=1…p) calculates the summation of its weighted input values $z_{inj} = v_{oj} + \sum_{i=1}^{n} x_i v_{ij}$, then apply the activation function $Z_j = f(z_{inj})$ and directs this value to all the output units.
6) Each output unit (y_k, k=1… m) find the summation of its weighted input values $y_{ink} = w_{ok} + \sum_{j=1}^{p} z_j w_{jk}$, and applies the activation function for calculating the output signals$Y_k = f(y_{ink})$.
7) Each output unit (y_k, k=1… m) obtains a known (target) feature vector corresponding to an input feature vector. Error term is calculated as $\delta_k = (t_k - y_k) f(y_{ink})$.
8) Each hidden unit (z_j, j=1… n) finds the summation of the delta inputs from the units of the layer just above it $\delta_{inj} = \sum_{k=1}^{m} \delta_j w_{jk}$. The error value is computed as $\delta_j = \delta_{inj} \; f(z_{inj})$.
9) Each output unit (y_k , k=1… m) modifies its bias and weights (j = 0… p) the weight correction term is calculated as $\triangle w_{jk} = \alpha \delta_k z_k$ and the bias correction term is calculated as $\triangle W_{ok} = \alpha \delta_k$. Therefore, the following equations are obtained
W_{jk} (new) = W_{jk}(old) +$\triangle W_{jk}$,
W_{ok} (new) =W_{ok} +$\triangle W_{ok}$.
Each hidden neuron $(z_j, j = 1, \ldots p)$ modifies its weights and bias (i=0…n). The weight correction value is given by
$\triangle V_{ij} = \alpha \delta_j x_i$.
The bias correction value is given by
$\triangle V_{oj} = \alpha \delta_j$.
V_{jk}(new) = V_{jk}(old) +$\triangle V_{ij}$, V_{oj}(new) =V_{oj}+$\triangle V_{oj}$.
10)Test the stopping condition.

Steps 1 to 3 initializes the weights, steps 4-6 are called feed forward steps, steps 7-8 are called BP steps, step 9 updates weight and biases, and finally step 10 is stopping condition which may be the minimization of the errors, number of epochs etc.

4. Conclusion

In this paper a classification model for detecting breast cancer based on Back-Propagation Neural Network (BPNN) and PCA data pre-processing is proposed. Noisy or irrelevant features can be removed by using dimension reduction step of PCA. Classification time is efficiently reduced by dimension reduction method. The BP algorithm and mathematical formula presented here can be applied to any neural network. The computing time is reduced if the weights chosen are small at the beginning.

References

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.
[2] Research Activities January 2012[online]. Available: http://www.ahrq.gov/research/jan12/0112RA20.htm.
[3] Breast Cancer [online]. Available: http://www.womenshealth.gov/breast-cancer/what-is-breast-cancer.
[4] R. W. Sembiring and J. M. Zain "The Design of Pre-Processing Multidimensional Data Based on Component Analysis", Comput. and Inform. Sci., vol. 4, no. 3, pp. 106-115, May 2011.
[5] S.N. Sivanandam and S.N. Deepa, Principles of Soft Computing, 1st Indian ed. Wiley publication, 2008, pp. 64-65.
[6] A. Frank and A. Asuncion (2010). UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.
[7] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", Int. J. of Comput. and Inform. Technology, vol. 1, no. 1, pp. 2277 – 0764 , Sept. 2012.
[8] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", Proc. IS&T/ SPIE Int. Symp. on Electron. Imaging: Sci. and Technology, 1993, vol. 1905, pp. 861–870.
[9] W. H. Wolberg, W. N. Street, D. M. Heisey and O. L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine needle aspirates", Archieves of Surgery , vol. 130, no. 5, May 1995, pp. 511-516, doi: 10.1001/archsurg.1995.014300500 6101 0.

[10] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi- Classifiers", Int. J. Of Compute and Inform. Technology, vol. 1, no. 1, pp. 2277 – 0764, Sept. 2012.
[11] A. Ilin and T. Raiko, "Practical Approaches to Principal Component Analysis in the Presence of Missing Values", J. of Machine Learning Research , vol. 11, pp. 1957-2000, 2010.
[12] P. Heermann and N. Khazenie,"Classifcation of multispectral remote sensing data using a back-propagation neural network", IEEE Trans. on Geoscience and Remote Sensing, vol. 30, pp. 81-88 , 1992.
[13] Y. H. Pao, Adaptive Pattern Recognition and Neural Network, Addison-Wesley Publishing Company, 1989.
[14] T. Kiyan and T Yildirim. "Breast cancer diagnosis using statistical neural networks", J. of Elect. & Electron. Eng., vol .4-2, 2004, pp. 1149- 1153.
[15] M. M. Beg and M. Jain, "An Analysis Of The Methods Employed For Breast Cancer Diagnosis", Int. J. of Research in Comput. Sci., vol. 2, no. 3, 2012, pp. 25-29.
[16] F. Paulin and A. Santhakumaran, "Classification of breast cancer by comparing Back propagation training algorithms", Int. J. on Comput. Sci. and Eng., vol. 3, no. 1, Jan. 2011, pp. 327-332.
[17] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.1, January 2000.
[18] M. Dash and H. Liu, "Dimensionality Reduction", Wiley Encyclopedia of Computer Science and Engineering, 2008.
[19] P. Cunningham, "Dimension Reduction", August 2007.
[20] C. Baneriee, S. Paul, and M. Ghoshal, "A comparative study of different ensemble learning techniques using Wisconsin breast cancer dataset," in 2017 International Conference on Computer, Electrical and Communication Engineering, ICCECE 2017, 2018.
[21] S. Sayed, S. Ahmed, and R. Poonia, "Holo entropy enabled decision tree classifier for breast cancer diagnosis using Wisconsin (prognostic) data set," in Proceedings - 7th International Conference on Communication Systems and Network Technologies, CSNT 2017, 2018.
[22] O. N. Oyelade, A. A. Obiniyi, S. B. Junaidu, and S. A. Adewuyi, "ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets," Informatics Med. Unlocked, 2018.
[23] L. Jin et al., "Modified single-output Chebyshev-polynomial feedforward neural network aided with subset method for classification of breast cancer," Neurocomputing, 2019.H.H. Crokell, "Specialization and International Competitiveness," in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

Author Profile

Shweta Saxena received B.E degree in Information Technology and M. Tech degree (Hons.) in Computer Science & Engineering from Rajiv Gandhi Proudyogiki Voshwavidhyalaya, Bhopal in 2006 and 2013 respectively. She is presently pursuing Ph. D. under the guidance of Dr. Manasi Gyanchandani in Maulana Azad National Institute of Technology, Bhopal.

Mansi Gyanchandani received Ph.D. degree from Maulana Azad National Institute of Technology (MANIT), Bhopal. She is currently working as Assistant Professor in department of CSE, MANIT Bhopal. Her teaching experience is 20 years. She has guided several master's level dissertations and presently guiding Ph. D. level dissertations. She has various research publications in reputed International Journals and conference proceedings. She is Life member of ISTE. Her areas of Specialization are Big data, Pattern recognition, data mining, machine learning, Privacy Preservation, Artificial Intelligence, Expert System, Neural Networks, Intrusion Detection & Information Retrieval.