



Early detection and diet prediction for breast cancer using Ant Colonized Back Propagation Algorithm

K.Geetha,
Department of Computer Science,
Bharathidasan University,
Trichy-23, Tamil Nadu, INDIA

Dr.M.Manimekalai,
Director, Department of Computer Science,
Shrimati Indira Gandhi College,
Trichy-02, Tamil Nadu, INDIA

Abstract: *Breast cancer façade stern menace to the lives of people today. The second leading cause for death in women today is breast cancer. Early deduction of breast cancer is essential for reduce the loss of life. The dataset is collected from Wisconsin Breast Cancer database. This paper proposes a hybrid algorithm which combines the Ant Colony optimization Algorithm with Back propagation learning to produce new algorithm named Ant Colonized Back Propagation Algorithm (ACBPA). The results obtained have also proved that the proposed algorithm can classify and detect the breast cancer. After detecting whether the cancer is benign or malignant, the diet required for the particular cancer is suggested.*

1 Introduction

WHO has estimated that cancer has been the cause of death for millions of people all over the world. It has also been estimated that it has been increasing upto 50% for developing countries. Cancer has been defined by the American Cancer society (2008) as the generic term for a large group of diseases that can affect any part of the body. Breast cancer always affects the inner lining of milk ducts or the lobules that supply the ducts with milk [1]. The risk factors which cause breast cancer are classified as modifiable and non-modifiable. The risk for breast cancer is most likely among the women and in older age. The other risk factors include family history of breast cancer, age of first birth, age of menopause, body weight, alcohol consumption, exposure to radiation [2], higher hormonal levels and diet[3]. Smoking tobacco increase [4] the risk of breast cancer as the long term smokers has an increased risk of cancer [5].

Early detection of breast cancer can facilitate to decrease the likelihood of full growth of tumours. The ways to detect the breast cancer is either by clinical examination or self breast examination and mammography. It is required that the women above 40 years should go for clinical examination for once a year and those who between 20 to 40 are advised to examine every 3 years. Diagnosis is the practice of predicting the existence of breast cancer as either benign or malignant.

Classification is a data mining technique which involves the use of supervised machine learning techniques which assigns labels or classes to different objects and groups. It involves the process of model construction (analysis of training data for patterns) and model usage where the constructed model is used for classification. Classification accuracy is usually estimated as the percentage of test samples that are correctly classified. The paper is further organised as follows: Related works sections describes the literature studied the section 3 gives detail about the data set used and the proposed methodology for detection and diet prediction for breast cancer and discusses the result obtained and the paper is concluded with section 4.

II. Related Works

A number of papers have been documented and published on the use of data mining techniques in the classification of breast cancer risks.

According to Rajesh et al (6) who used SEER dataset for the diagnosis of breast cancer using the C4.5 classification algorithm. The algorithm was used to classify patients into either pre-cancer stage or potential breast cancer cases. Random tests were performed on the dataset which contained information for 1183 patients including the age of diagnosis, regional lymph nodes measures, and sequence number of tumors, dimension of primary tumor and contiguous growth of the primary tumor. The analysis involved the use of three random 500 records from the pre-processed data of 1183 and was used as training data and the lowest error rate achieved was 0.599. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.

Shajahan et al (7) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10

attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes' had the an accuracy of 97.42% with an error rate of 0.0258.

Mangasarian et al (8) performed classification on both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method-Tree (MSM-T) that uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM-T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) that uses linear programming to determine a linear combination of the input features which accurately predicts the Time-To-Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM-T approach was 97.3% and 97 % respectively whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

Lundin et al (9) has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku. To evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. From the experiment the values of ROC curve for 5 years was evaluated as 0.909, for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. The author compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model.

Delen et al (10) compared ANN, decision tree and logistic regression techniques for breast cancer prediction analysis. They used the SEER data of twenty variables in the prediction models. From the experiment the author found that the decision tree with 93.6% accuracy and ANN with 91.2% are more superior to logistic regression with 89.2% accuracy. The study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory.

III. Ant Colonized Back Propagation Algorithm (ACBPA)

The data required for cancer prediction is collected from Wisconsin Breast Cancer Database. [11] . It has a total of 699 instances. The number of attributes taken is 10. The list of total attributes is given in Table 1.

Attribute	Domain
Clump Thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal Adhesion	1-10
Single Epithelial cell size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	2 for benign , 4 for malignant

Table 1: Description of the attributes

The proposed Algorithm combines Ant Colony Optimization Algorithm with Back propagation Algorithm. The two algorithm was taken to hybrid with each other since they both can back propagate and re check and produce the result required. The proposed Algorithm is as follows:

Ant Colonized Back Propagation Algorithm:

Input: 9 Attributes

Algorithm:

Step 1: Start

Step 2: Initialize pheromone, heuristics, probabilities of edges and connection weights.

Step 3: Create Input Vector $a_i = [a_{i1} \ a_{i2} \ a_{i3} \ a_{i4} \ a_{i5} \ a_{i6} \ a_{i7} \ a_{i8} \ a_{i9}]$ as ants.

Create Output target $O_{ji} = [O_{j1}, O_{j2}]$

Step 4: Let ant run from input to output layer.

Step 5: For every input node I in every layer

$$\gamma_{op} = \alpha_{ip} \dots (1)$$

Step 6: For every input/ant p in every layer, $j=1,2,3 \dots M$ from input to output layer, find the output.

$$\gamma_{jp} = f \left(\sum_{k=1}^{Nj-1} \gamma_{(j-1)k} W_{jpk} \right) \dots (2)$$

Where

$$f(x) = \frac{1}{1 + \exp(x)} \dots (3)$$

Step 7: Obtain output value. For every output node in layer M , perform

$$O_{pi} = \gamma_{Mi} \dots (4)$$

Step 8: Evaporate the pheromone.

Step 9: Update the path.

Step 10: Adjust pheromone level

Step 11: Update probability of edge.

Step 12: Calculate the error

$$\delta_{Mi} = \gamma_{Mi} (1 - \gamma_{Mi}) (O_{ji} - \gamma_{Mi}) \dots (5)$$

Step 13: If error is not minimized,

Goto Step 2

Step 14: End

Output:

Risk prediction

The parameters such as Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE), True Positive Rate (TPR), False Positive Rate (FPR), Regression, Mean Squared Error (MSE) are used to evaluate the performance of the proposed Algorithm. The data set is experimented with Ant Colony Optimization Algorithm and Back propagation Learning Algorithm. The results obtained are tabulated as in Table 2. When comparing the results obtained it can be proved that the proposed algorithm is better than the existing algorithm in classifying the breast cancer as benign and malignant.

Parameters	ACO	BPL (NN)	ACBPA
Accuracy	95.5651	96.2804	98.1402
RMSE	0.1958	0.1889	0.1213
RRSE	41.1962	39.7371	18.5623
MSE	5.65609e-2	3.36426e-2	1.55368e-2
TPR	0.956	0.963	0.987
FPR	0.055	0.037	0.023
Regression	8.80319e-1	9.30442e-1	9.69137e-1

Table 2: Comparison of results obtained

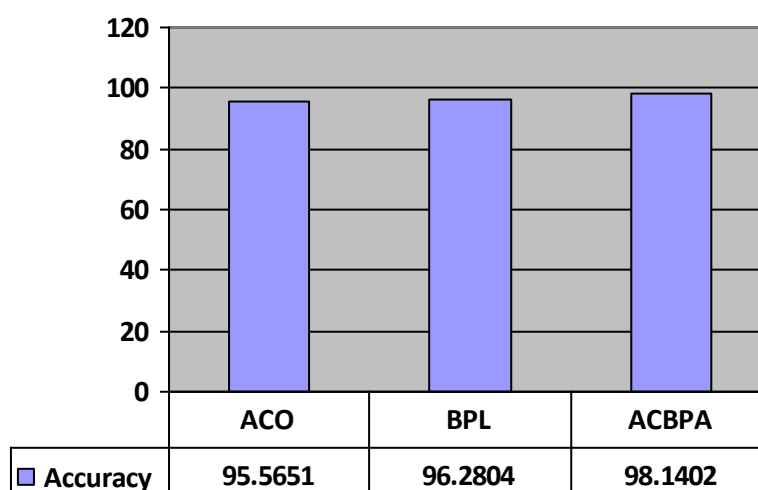


Figure 1: Comparison of accuracy (%)

Figure 1 explain the prediction accuracy level obtained while using each algorithm. The proposed algorithm has given an accuracy level of 98.14% which is better than the other two algorithms.

ACO			BPL			ACBPA		
	a	b		a	b		a	b
a	441	17	a	443	15	a	451	6
b	9	232	b	16	225	b	7	234

Table 3: confusion Matrix where a=benign, b=malignant

The table 3 explains the confusion matrix obtained. In ACBPA, the correctly classified instance is much greater than the other two. Moreover, the error possibility is also reduced in case of ACBPA.

IV. Conclusion

In this paper, two different algorithms are experimented and a new hybrid algorithm which combines the important features of the existing algorithm is proposed. The proposed ACBPA performs well in predicting the cancer. An accuracy of 98.14% is obtained using the ACBPA. After classifying the patients, the required diet for the patients can be suggested with a help of medical practitioner. This algorithm can be developed as a software tool which will help the medical people in easily diagnosing the cancer. This is the future direction of this work.

References

- [1]. Sariego J (2010). "Breast cancer in the young patient". *The American surgeon* **76** (12): 1397–1401. PMID 21265355.
- [2]. Poongodi, M., Manjula, L., Pradeepkumar, S., Umadevi, M. *Cancer Prediction Technique Using Fuzzy Logic*. International Journal of Current Research, Vol. 3, Issue 11, pg 333-336, December 12, 2001. <http://www.journalera.com>. ISSN: 0975-833X. Accessed on June 24, 2014.
- [3]. Yager JD (2006). "Estrogen carcinogenesis in breast cancer". *New Engl J Med* **354** (3): 270–82. doi:10.1056/NEJMra050776. PMID 16421368.
- [4]. Johnson KC, Miller, AB, Collishaw, NE, Palmer, JR, Hammond, SK, Salmon, AG, Cantor, KP, Miller, MD, Boyd, NF, Millar, J, Turcotte, F (2009). "Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009)". *Tobacco control* **20** (1): e2. doi:10.1136/tc.2010.035931. PMID 21148114.
- [5]. Santoro, E., DeSoto, M., and Hong Lee, J (February 2009). "Hormone Therapy and Menopause". National Research Center for Women & Families. <http://www.center4research.org/2010/03/hormone-therapy-and-menopause/>.
- [6]. Rajesh, K., Anand, S (2012). Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012. ISSN 2278-1021. <http://www.ijarcee.com> pg. 72 – 77.
- [7]. Shajahaan, S.S; Shanthi, S., Chitra, V.M. (2013). Application of Data Mining Techniques to model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering Vol 3, Issue 11, November 2013. ISSN 2250-2459. <http://www.ijetac.com> pg 362 – 369.
- [8]. Mangasarian,D.S.;Street, W.N.,Wolberg, W.H (1995). Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), pages 570-577, July-August 1995.
- [9]. Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pytkänen L. and Joensuu H.,(1999) "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.
- [10]. Delen, D., Walker, G., Kadam, A. (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, pp. 113-127, June 2005.
- [11]. [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))