A Novel Approach for Predicting Breast Cancer using Data Mining Techniques

[1] S. AMARESAN, [2]J. MARIYA GOLDA,

[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2] Prist University, Thanjavur.

Abstract : Women who have improved from breast cancer (BC) constantly panic about setback. The way that they have persevered through the meticulous treatment makes repeat their biggest fear. However, with current spreads in technology, early repeat prediction can enable patients to get treatment prior. The accessibility of broad information and propelled techniques make precise and fast prediction possible. This examination expects to think about the exactness of a couple of existing information mining calculations in predicting BC repeat. It inserts a particle swarm optimization as highlight choice into ANN classifier. An objective of increasing the accuracy level of the prediction model**.**

***IndexTerms* - Particle Swarm Optimization, Naïve Bayes, Artificial Neural Network, K-Nearest Neighbor, Feature Selection, Classification, Wiscon breast cancer dataset.**

I. Introduction

Breast cancer is the most frequently diagnosed cancer and the leading cause of death among females, accounting for 23% of the total cancer cases and 14% of cancer deaths. It is by far the most commonly diagnosed cancer in women. The global burden of cancer continues to increase largely because of the aging and growth of the world population alongside an increasing adoption of cancer-causing behaviours, particularly smoking, in economically developing countries.

It is also estimated that by 2030 the global burden of breast cancer will increase to over two million new cases per year. Furthermore, it is estimated that this increase in cases will be largely due to increasing incidence in developing regions of the world. Mammograms are the best way to screen the breast cancer at an early stage that is, before it can be felt, it is easier to treat. The Figure 1.1 shows that the mortality rates of breast cancer were estimated per 1, 00, 000 women worldwide.

Organ chlorines are considered a possible cause for hormone-dependent cancers (Siddiqui et al., 2004). Detection of early and subtle signs of breast cancer requires high-quality images and skilled mammographic interpretation. Radiologists should be trained in the recognition of the signs of early onset of reading mammograms, which may be subtle and may not show typical malignant features.

A recent report by the Indian Council of Medical Research predicted that the number of breast cancer cases in India to rise to 106,124 in 2015 and to 123,634 in 2020. The Figure 1.2 shows that the mortality rates of breast cancer were estimated per 1, 00,000 women in India.

The American Cancer Society (ACS) has recommended that women between the ages of 40 and 49 have a mammography examination every other year, and women over 50 have an examination every year. The risk of developing breast cancer can be reduced by: (a) Having children before 30, (b) Breast-feeding, (c) Limiting alcohol intake, (d) Maintaining a healthy weight, (e) Exercising regularly.

II. RELATED WORKS

Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari [R] The objective of this research paper is to present a report on breast cancer where we took advantage of those available technological advancements to develop prediction models for breast cancer survivability. The authors used three popular data mining algorithms (Naïve Bayes, RBF Network, J48) to develop the prediction models using a large dataset (683 breast cancer cases). The authors also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The results (based on average accuracy Breast Cancer dataset) indicated that the Naive Bayes is the best predictor with 97.36% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), RBF Network came out to be the second with 96.77% accuracy, J48 came out third with 93.41% accuracy.

Shaikh, Tawseef Ayoub, and Rashid Ali [r] This paper uses dimensionality reduction technique offered by Weka tool called WrapperSubsetEval on two benchmark cancer datasets of Wisconsin and Portuguese “Breast Cancer Digital Repository” (BCDR), on top four data mining algorithms available in literature. The final experiments carried in MATLAB andWeka demonstrated that Naive Bayes, J48, k-NN and SVM got an improvement in accuracy from 92.6186, 92.9701, 96.1336, 97.891 to 97.0123, 96.8366, 97.3638, 97.9123% in case of Wisconsin dataset and an improvement from 87.4126, 80.4196, 93.7063, 91.6084 to 89.5105, 90.9091, 97.9021, 95.1049% in case of BCDR-D01_Dataset.

Sri, M. Navya, et al [R] In classification, in order to develop a model which will categorize the population of records, the authors make use of a set of pre-classified examples. The techniques of classification use the model which is built on basis of training data and apply it to test data. “Breast cancer Wisconsin data set is used as a training set.” There is an open source data mining tool named WEKA, which consists of implementation of data mining algorithms. By making use of WEKA the authors have compared the well-known classification algorithms that are decision tree and Bayesian algorithms. It is concluded that decision tree classification algorithm got high accuracy compared to Bayesian classification algorithm.

Dutta, Soumi, et al [R] The major scope of this research work is to predict the possibilities of breast cancer based on existing clinical records. Fuzzy inference-based classification system is proposed in this work to predict unseen instances. The tumors genetic behaviors are analysed for prediction modeling.

Borah, Rupam, Sunil Dhimal, and Kalpana Sharma [R] This project aims to demonstrate the working and the accuracy of a few machine learning models on a given set of data on breast cancer and also, making a comparison between them to determine the best model suitable of a particular paradigm.

III. PROPOSED FRAMEWORK FOR CLASSIFICATION OF BREAST CANCER

Figure 1 depicts the proposed framework for the classification of breast cancer.

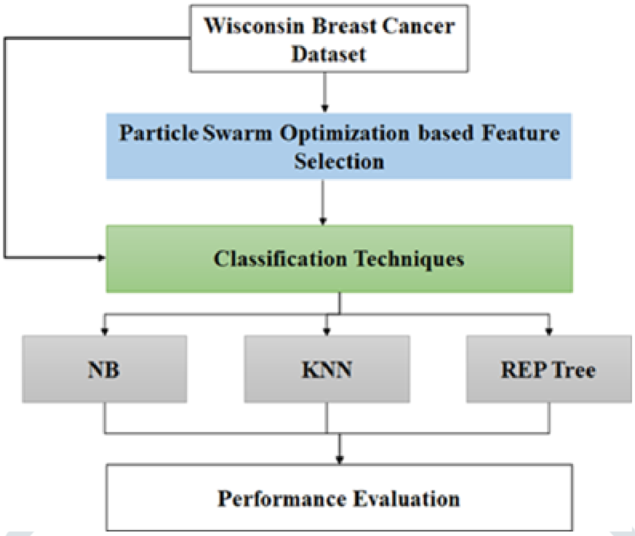
Wisconsin Breast Cancer Dataset
Particle Swarm Optimization based Feature Selection
Classification Techniques
NB
KNN
REP Tree
Performance Evaluation

Figure 1: Proposed framework for the classification of breast cancer

3.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is based on the social behavior associated with bird's flocking for optimization problem. A social behavior pattern of organisms that live and interact within large groups is the inspiration for PSO. The PSO is easier to lay into operation than Genetic Algorithm. It is for the motivation that PSO doesn't have mutation or crossover operators and movement of particles is affected by using velocity function. In PSO, every particle alters its own flying memory and its partner's flying inclusion keeping in mind the end goal to flying in the search space with velocity [17][18][19].

The best-fit particle of the entire swarm influences the position of each particle. Each individual particle j ∈ $[1 \ldots m]$ where m > 1, has current position in search space s_j, a current velocity u_j and a personal best position $p_{b,j}$ where j is the smallest value determined by objective function o. By using $p_{b,j}$ the global best position G_b is calculated, which is the buck value obtained by comparing all the $p_{b,j}$

The $p_{b,j}$ is calculated by using the formula

$$p_{b,j}= \begin{cases} p_{b,j} \; if \; (y_j) > \; p_{b,j} \\ y_j \; if \; f(y_j) \; \leq p_{b,j} \end{cases}$$

The formula used to calculate Global Best Position G_{best} is

$$G_b \; = \; \{\min\{p_{b,j}\}, where \; j \; \in [1, \ldots.., m] where \; m > 1$$

Velocity can be updated by using the formula

$$u_j^{j+1} \; = \; wu_j(t) + \; s_1 i_1[y_j\,(t) - y_j\,(t)] + \; d_2 i_2[g(t) - \; y_j\,(t)\,]$$

where $u_i(t)$ is the velocity and w, s_1 and s_2 are used supplied co-efficient. The i_1 and i_2 are random values $y_j(t)$ is the individual best solution, g(t) is the swarm's global best candidate solution. $wu_j(t)$ is known as inertia component. Inertia component value lies between 0.8 and 1.2. Lower the values of inertia component, it speeds up the convergence of swarm to optima. But higher value encourages the exploration of entire search space. $s_1 i_1[y_j\,(t) - y_j\,(t)]$ is known as cognitive component.

The following steps are done in PSO algorithm:

```
1. Initialize each particle in the population with random positions and velocities.
2. Repeat the following steps until stopping criterion is met.
        i. for each particle
        {
                Calculate the fitness function value;
                Compare the fitness value:
                If it is superior to the best fitness value pbest, then current value is assigned pbest
        value;
        }
        ii. Best fitness value particles among all the particles are selected and assign it as gbest;
        iii. for each particle
        {
                Calculate particle velocity;
                Change the position of the particle;
        }
```

3.2 Naïve Bayes Classification

Naïve Bayesian classifier is a simple classification scheme, which estimates the class- conditional probability by assuming that the attributes are conditionally independent, given the class label. Naive Bayes is a strategy for assessing probabilities of individual variable qualities, given a class, from preparing information and to then permit the utilization of these probabilities to order new elements, which is a term in Bayesian insights managing a straightforward probabilistic classifier taking into account applying Bayes' hypothesis (from Bayesian measurements) with strong (guileless) autonomy assumptions. In basic terms, a strong Bayes classifier expect that the nearness (or nonappearance) of a specific feature of a class is disconnected to the nearness (or nonattendance) of some other element. The Naive Bayesian classifier, fills in as taking after inference [20][21][22][23][24]:

Step 1: Let T be a training set of tuples and their related class names. Each tuple is spoken to by a m-dimensional attribute vector, A = (a1, a2, … .. , am), m estimations made on the tuple from m properties, individually, X1, X2, … , Xm.

Step 2: Suppose that there are n classes D1, D2, … . , and Dn. Given a tuple, A, the classifier will anticipate that A has a place with the class having the most noteworthy back likelihood, adapted on A. That is, the guileless Bayesian classifier predicts that tuple A has a place with the class Tj if and just if

$$P\ ((D_j|A) > P\ ((D_k|A)\ ) \text{ for } 1 \leq k \leq n,\ k\ \neq j \qquad (7)$$

The boost P(Dj|A). The class Dj for which P(Dk|A) is amplified is known as the most extreme posterior hypothesis. By Bayes' hypothesis (Next condition)

$$P(D_j|A) = \frac{P(A|D_j)P(D_j)}{P(A)} \qquad (8)$$

Step 3: Since P(A) is consistent for all classes, just (P(Dj|A) = P(A |Dj)P(Dj)) should be amplified.

Step 4: Based on the supposition is that properties are restrictively free (i.e., no reliance connection between attributes), the registering of P(A|Dj) utilizing the accompanying condition:

$$P(A|D_j) = \prod_{i=1}^{m} P(a_i|D_j) \qquad (9)$$

Diminishes the calculation cost by Equation (P(Dj|A) = P(A |Dj)P(Dj), just numbers the class appropriation. On the off chance that Xi is unmitigated, P(Ai|Dj) is the no. of tuples in Dj having esteem Ai for Xi separated by |Dj, T| no. of tuples of Dj in T. Also, if Xi is persistent esteemed, P(Ai|Dj) is typically processed in view of Gaussian circulation with a mean μ and standard deviation σ and P(Ai|Dj) is:

$$g\ (x,\ \mu,\ \sigma\ ) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (10)$$

$$P(D_j|A) = g\ (a_i, \mu_{Dj}, \sigma_{Dj}) \qquad (11)$$

Where μ is the mean and σ is the difference. On the off chance that a property estimation doesn't happen with each class esteem, the likelihood will be zero, and a posteriori likelihood will likewise be zero.

3.3 KNN Classification

The k-nearest neighbors algorithm is one of the most used algorithms in machine learning. It is a learning method bases on instances that does not required a learning phase. The training sample, associated with a distance function and the choice function of the class based on the classes of nearest neighbors is the model developed. Before classifying a new element, we must compare it to other elements using a similarity measure. Its k-nearest neighbors are then considered, the class that appears most among the neighbors is assigned to the element to be classified.

3.4 REP Tree Classification Method

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance.

IV. RESULT AND DISCUSSION

4.1 Dataset Description

Table 1 depicts the description of the Wisconsin breast cancer Dataset.

Table 1: Description of the Wisconsin Breast Cancer Dataset

Feature Number	Feature Name	Description of the Feature
1	Id	ID number of the patient
2	radius_mean	mean of distances from center to points on the perimeter
3	texture_mean	standard deviation of gray-scale values
4	perimeter_mean	mean size of the core tumor
5	Area_mean	Mean of the tumor area
6	smoothness_mean	mean of local variation in radius lengths
7	compactness_mean	mean of perimeter^2 / area - 1.0
8	concavity_mean	mean of severity of concave portions of the contour
9	concave points_mean	mean for number of concave portions of the contour
10	symmetry_mean	
11	fractal_dimension_mean	mean for "coastline approximation" – 1
12	radius_se	standard error for the mean of distances from center to points on the perimeter
13	texture_se	standard error for standard deviation of gray-scale values
14	perimeter_se	
15	Area_se	
16	smoothness_se	standard error for local variation in radius lengths
17	compactness_se	standard error for perimeter^2 / area - 1.0
18	concavity_se	standard error for severity of concave portions of the contour

19	concave points_se	standard error for number of concave portions of the contour
20	symmetry_se	
21	fractal_dimension_se	standard error for "coastline approximation" – 1
22	radius_worst	"worst" or largest mean value for mean of distances from center to points on the perimeter
23	texture_worst	"worst" or largest mean value for standard deviation of gray-scale values
24	perimeter_worst	
25	area_worst	
26	smoothness_worst	"worst" or largest mean value for local variation in radius lengths
27	compactness_worst	"worst" or largest mean value for perimeter^2 / area - 1.0
28	concavity_worst	"worst" or largest mean value for severity of concave portions of the contour
29	concave points_worst	"worst" or largest mean value for number of concave portions of the contour
30	symmetry_worst	
31	fractal_dimension_worst	"worst" or largest mean value for "coastline approximation" – 1
32	Diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)

The following table 2 depicts the number of features obtained by using Particle Swarm Optimization technique feature selection. The following 8 features are obtained by PSO search space with Correlation feature selection (CFS) method.

Table 2: Number of Features obtained by PSO based Feature Selection Method

Sl.No	Name of the Feature
1	perimeter_mean
2	area_mean
3	concavity_mean
4	concave points_mean
5	concavity_se
6	radius_worst
7	area_worst
8	concave points_worst

The following classification techniques like Naïve Bayes, K-Nearest Neighbor and REP Tree are used to check the classification accuracy and error rates of the Breast Cancer dataset without using feature selection method.

Table 3: Performance analysis of the original dataset without feature selection using NB, KNN and REP Tree classificatio**n** **methods**

Performance Metric	**Classification Techniques**		
	Naïve Bayes	**K-Nearest Neighbor**	**REP Tree**
Accuracy	43.058 %	74.8682 %	62.7417 %
Kappa Statistic	0.0692	0.4371	0
Mean absolute error	0.5438	0.2513	0.4675
Root mean squared error	0.7002	0.5013	0.4835
Relative absolute error	116.2686 %	53.7387 %	99.973%
Root relative squared error	144.8263 %	103.6839 %	100%
TP Rate	0.431	0.749	0.627
FP Rate	0.431	0.331	0.627
Precision	0.757	0.744	0.627
Recall	0.431	0.749	0.627
F-Measure	0.320	0.741	0.771
ROC Area	0.789	0.709	0.493
PRC Area	0.793	0.676	0.529

The following classification techniques like Naïve Bayes, K-Nearest Neighbor and REP Tree are used to check the classification accuracy and error rates of the Breast Cancer dataset with using feature selection method. Table 4 depicts the performance analysis of the reduced dataset obtained by PSO feature selection method.

Table 4: Performance analysis of the reduced dataset with feature selection using NB, KNN and REP Tree classification methods

Performance Metric	**Classification Techniques**			
	Naïve Bayes	**K-Nearest Neighbor**	**REP Tree**	**ANN**
Accuracy	94.2004 %	93.8489 %	93.1459 %	**95.9578 %**
Kappa Statistic	0.8756	0.8688	0.8533	**0.913**
Mean absolute error	0.063	**0.0615**	0.0913	0.0702
Root mean squared error	0.2322	0.248	0.241	**0.1794**
Relative absolute error	13.4632 %	**13.1528 %**	19.5177 %	15.0039 %
Root relative squared error	48.0259 %	51.2953 %	49.8402 %	**37.1078 %**

TP Rate	0.942	0.938	0.931	**0.960**
FP Rate	0.069	0.067	0.079	**0.053**
Precision	0.942	0.939	0.931	**0.960**
Recall	0.942	0.938	0.931	**0.960**
F-Measure	0.942	0.939	0.931	**0.959**
ROC Area	0.982	0.936	0.956	**0.986**
PRC Area	0.982	0.913	0.945	**0.987**

From the table 4, Artificial Neural Network classification technique performs better than the other classification techniques like KNN, NB and REP Tree. The accuracy, TP, Precision, Recall, F-Measure, Kappa Statistic, ROC area, PRC area are increased when reduced dataset is executed with ANN classifier than the other classifiers. Root Mean Squared Error (RMSE) and Root Relative Squared Error (RRSE) are reduced using ANN classification.

Artificial Neural Network classification technique gives better classification accuracy when using reduced dataset with PSO based feature selection method.

V. CONCLUSION

This project has focused on the investigating the effect of integrating the feature selection algorithm with classification algorithms in the detection of breast cancer. The performance of the classification method can be improved by using feature selection techniques to reduce the number of features (considering most relevant, removing irrelevant, and redundant features). Some features have more influence and importance over the results of the classification algorithms compared to other features. The most popular three classification techniques like Naïve Bayes, KNN and REP tree are considered for evaluating the original dataset ie without using feature selection method. In this project, proposed breast cancer detection framework with PSO based feature selection and ANN classification performs better than feature selection with other classification techniques. ANN with feature selection gives increased accuracy, True positive rate, Precision, Recall, F-measure, ROC area, Kappa Statistic and PRC area. Error rates like RMSE and RRSE are reduced when using ANN with PSO based feature selection method.

The future direction of this study will include testing newer algorithms with other feature selection techniques. We will experiment on cluster techniques as well as ensemble algorithms.

REFERENCES

[1] GLOBOCAN Cancer Fact Sheets: Breast Cancer, [Online] GLOBOCAN 2008 (IARC) http://globocan.iarc.fr/factsheets/cancers/breast.asp,2008.

[2] American Cancer Society. Detailed Guide: Breast Cancer [Online] http://www.cancer.org, 2012.

[3] Siddiqui, M. K. J., Anand, M., Mehrotr, P.K., Sarangi, R., and Mathur, N., “Biomonitoring of Organochlorines in Women with Benign and Malignant Breast Disease”, *Environmental Research*, Vol. 1, pp. 1-8, 2004.

[4] Development of an Atlas of Cancer in India: National Cancer Registry Programme (Indian Council of Medical Research) [Online] http://www.ncrpindia.org/Cancer_Atlas_India/contactus.aspx, 2010.

[5] Moinfar, F (2007). *Essentials of Diagnostic Breast Pathology*. Springer.

[6] Moore K. L, Agur A.M and Dalley A.F (2004). *Essential Clinical Anatomy*. (4nd ed.). (W. Kluwer, Ed.).

[7] Kronberger, L., Steinschifter, W., Weblacher, M., Estelberger, W., Liebmann, P. M., Rabl, H., Smola, M., Lax, S. F., Mischinger, H. J., Schauenstein, E., and Schauenstein, K., “Selective decrease of serum immunoglobulin G1 as marker for early stages of invasive breast cancer”, Breast Cancer Research and Treatment, Kluwer Academic Publishers, Vol. 64, pp. 193-199. 2000.

[8] Gray H (2000). *Anatomy of the Human Body*. New York: Bartleby.

[9] Dixon J.M (2006). *ABC of Breast Diseases*. (3rd Ed.) B. Publishing, Ed.

[10] Seeley R, Stephens T and Tate P (2004). Anatomy and Physiology. The McGraw-Hill Company.

[11] Guyton A.C and Hall J.E (2000). *Textbook of Medical Physiology* (10th Ed). (W. S. Company, Ed.

[12] Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." *Journal of Algorithms & Computational Technology* 12.2 (2018): 119-126.

[13] Shaikh, Tawseef Ayoub, and Rashid Ali. "Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk." *Proceedings of 2nd International Conference on Communication, Computing and Networking*. Springer, Singapore, 2019.

[14] Sri, M. Navya, et al. "A Comparative Analysis of Breast Cancer Data Set Using Different Classification Methods." *Smart Intelligent Computing and Applications*. Springer, Singapore, 2019. 175-181.

[15] Dutta, Soumi, et al. "Cancer Prediction Based on Fuzzy Inference System." *Smart Innovations in Communication and Computational Sciences*. Springer, Singapore, 2019. 127-136.

[16] Borah, Rupam, Sunil Dhimal, and Kalpana Sharma. "Medical Diagnostic Models an Implementation of Machine Learning Techniques for Diagnosis in Breast Cancer Patients." *Advanced Computational and Communication Paradigms*. Springer, Singapore, 2018. 395-405.

[17] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.

[18] M. Durairaj, T S Poornappriya, “Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.”, *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[19] M. Durairaj, T. S. Poornappriya, “Importance of MapReduce for Big Data Applications: A Survey”, *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.

[20] M. Lalli, V.Palanisamy,(2016), “Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks”, International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[21] M. Lalli, V.Palanisamy,(2017), “Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques”, Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[22] M. Lalli, V.Palanisamy, (Sep 2014), “A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN”, International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[23] M. Lalli, “Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET”, *Journal of Applied Science and Computations,* Volume VI, Issue VI, JUNE/2019, 1795-1813.

[24] M. Lalli, “Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network”, *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.