Development of Machine Learning Model Using Least Square-Support Vector Machine, Differential Evolution And Back Propagation Neural Network to Detect Breast Cancer

Madhura D. Vankar[1], Prof. G. A. Patil[2]
[1]Dept of Computer Science And Engineering
[2]H.O.D, Dept of Computer Science And Engineering
[1, 2] Shivaji University, D. Y. Patil College of Engineering and Technology,
KasabaBawada,Kolhapur, Maharashtra, India.

Abstract- *In developed countries, Breast Cancer has become a major cause of death among women. Breast Cancer diagnosis has been a machine learning problem. Various Machine learning algorithms have been widely used. The problem is basically nontrivial and not easy to solve because the dataset is relatively small and noisy. Here training dataset will be taken from UCI repository. Therefore System will compare result of LS-SVM and back propagation neural network to generate accurate result at early stage. Early diagnosis requires correct and reliable diagnosis procedure that will allow physicians to distinguish benign breast tumors from malignant ones. It will be helpful to medical field because it avoids loss of data.*

Keywords- Differential Evolution(DE), Least-square support vector machine(LS-SVM), Back-propagation neural network(BPN), Confusion matrix, cross validation, Features learning model.

I. INTRODUCTION

The use of machine learning tools in medical diagnosis is increasing gradually. This is mainly because the effectiveness of classification and recognition systems has improved in a great deal to help medical experts in diagnosing diseases. Such a disease is breast cancer, which is a very common type of cancer among woman. In this paper, breast cancer diagnosis was conducted using least square support vector machine (LS-SVM) classifier algorithm and back-propagation neural network. The robustness of the LS-SVM is examined using classification accuracy, analysis of sensitivity and specificity, *k*-fold cross-validation method and confusion matrix.

This paper proposes a hybrid classification algorithm using Differential Evolution (DE) and Least Squares Support Vector Machine (LS-SVM). LS-SVM technique is used for classification. Since LS-SVM classifier is so sensitive to the changes of its parameter values, DE algorithm is used as an optimization technique for LS-SVM parameters. This will guarantee the effectiveness of the hybrid algorithm by searching for the optimal values of the classifier. The aim of this paper is to help physicians in the early diagnosis for BC Patients.

The Least Square Support Vector Machine (LS-SVM) was first proposed by Suykens and et al. by modifying the formulation of standard SVM. The LS-SVM was modified at two points: First, instead of inequality constraints, it takes equality constraints and changed the quadratic programming to a linear programming. Second, a squared loss function is taken from the error variable. In this study, LS-SVM was employed to diagnose the breast cancer. For training and testing experiments, WBCD taken from the University of California at Irvine (UCI) machine learning repository was used. In this study, the performance was evaluated by the well-known k-fold cross validation method.

II. LITERATURE SURVEY

Md. KamrulHasan , Md. Milon Islam, and M. M. A. Hashem[1], have developed a 10 fold cross validated mathematical model by using symbolic regression of multigene genetic programming to detect breast cancer. It requires high training time.

Parveen, Amritpal Singh [2] have proposed hybrid methodology of combining support vector machine(SVM) and fuzzy c-means clustering for classification. It gives accurate result for identifying the brain tumour. It has high algorithm complexity and requires extensive memory.

Yu-Ling Hou, Chih-Min Lin, Kuo-HsinChen, Te-YuChen [3] have proposed breast nodule CAD system for characterizing breast nodules as either benign or malignant on an ultrasonic image. This paper described a fuzzy cerebellar model neural network (FCMNN) as a classifier. All analysis parameters are decided here on the basis of trial and error.
Satish Saini ,Ritu Vijay [4] have described Feed-forward back-propagation Artificial Neural Network (ANN) model for detection of breast cancer using Image Registration Techniques. The performance of the system was evaluated on the basis of Mean Square Error (MSE). In this paper, Number of neurons are computed on trial and error method, thereby it is consuming more time.

Chandra PrasetyoUtomo, AanKardiana, Rika Yuliwulandari [5] have proposed medical decision support systems using Extreme Learning Machine Neural Networks (ELM ANN). ELM is a tuning free algorithm that helps in achieving high sensitivity and accuracy rates .This paper has limitations such as low specificity rate and takes more time to train than other methods.

Seema Singh ,Sunita Saini, Mandeep Singh [6] have proposed model to detect cancer using adaptive neural network for detecting the cancer stage as benign or malignant. It performs clustering and allows the number of clusters to vary with the size of the problem. It has better accuracy than fuzzy system. It is a slow learning process.

III. PROBLEM DEFINITION

To Develop Machine Learning Model to optimize training time and enhance performance by using Least Square-Support Vector Machine, Differential Evolution and Back Propagation Neural network, so as to help detecting more accurately the breast cancer in the patients.

IV. OBJECTIVES

The main objectives of this research work are:

- To develop Machine Learning Model.
- To replace missing value in dataset using suitable algorithm.
- To optimize training time by Merging Differential Evolution with LS-SVM.
- To obtain Accurate detection results by comparing LS-SVM and BPN.
- To evaluate Performance of system with the help of 10-fold cross validation method.

V. PROPOSED SYSTEM ARCHITECTURE

Our proposed system will be suitable especially for big data machine learning with the help of training model. Dataset will be processed by using suitable missing value replacement method to improve system's performance. Data will be converted into binary format by using 'Feature Scaling'. After conversion, data will be transferred towards learning model to train neural network by using two classifiers such as LS-SVM and BPN. The System is expected to compare results of two classifiers and provide accurate result to end-user with the help of confusion matrix.

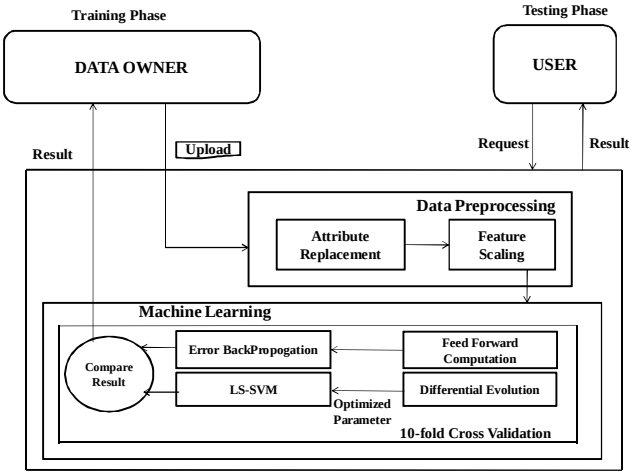

Fig. 1 Proposed System Architecture

VI. METHODOLOGY

1. Dataset Attributes:

The proposed algorithm will work on the Wisconsin Breast Cancer Dataset (WBCD) taken from UCI repository of machine learning databases**.** The UCI data consists of nine input variables and one output (2, 4). The numeric value 2 represents that it is of benign type and 4 represents that it is of malignant type of breast cancer. The database contains the 'sample code number' which is not required for classification process, which is discarded.

Table 1: Attributes to be considered from the WBCD

	Attributes	Domain
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	Mitoses	1 - 10
10	Class	2 for benign, 4 for malignant

The proposed system has been divided into following modules as-

Module 1: Missing value replacement and Feature Scaling of Attributes:

This module will be used to make dataset ready for training neural network using following methods.

1.1 Replacement of missing values within dataset using suitable algorithm:

Missing data or missing values occur when no data value is stored for an instance in the current record. Missing values may occur for several reasons and affects the quality of data, such as unavailability of data, malfunctioning of measurement equipment, changes in experimental design during data collection; sometimes data is not presented or get corrupted due to inconsistency of data files. To compute these missing values three techniques will be used named as Litwise deletion, Mean/mode imputation and KNN imputation, which result in imputed datasets.

A. Litwise deletion algorithm / Ignoring the tuples containing missing data:

This method consists of discarding all instances (cases) with missing values for at least one feature. The simplest way of handling missing data is to delete the subjects that have missing values. It is available in all statistical packages and is the default method in many programs. After applying litwise deletion all the cases having missing values are deleted from the dataset.

B. K-nearest neighbors algorithm1:

In this method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances will be determined by using a distance Function.

Distance function can be Euclidean and Manhattan etc. In this work we will considerthe Euclidean distance.

K-nearest neighbors algorithm is as follows-

i) Determine the value of K(Nearest neighbors).Value of K will be chosen randomly.
ii) Calculate the distance between the missing value instance and other training instance i.e. based upon the value of K.
iii) Euclidean distance is given by the equation as:-

$D(x,y) = \Sigma\sqrt{xi2 - yi2}$ ---------- (1)

This can be interpreted as below:

Dist(X, Y) = $\sqrt{(X-X1)2+(Y-Y1)2}$ ---------- (2)

C. Mean/mode imputation:

This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs.

Mean is calculated as:

Mean = Sum of all the value / Total no of values

1.2 Feature Scaling of Attribute:

Scaling has the advantage of mapping the desired range of variables ranging between minimum and maximum range of network input. The conversion of the given data sets into binary will be done based on certain ranges, which are defined for each attribute. First from the given range of inputs, the minimum and maximum value will be picked up.

This scaling will be done by the following formula:

X' = X- MIN(X) /MAX(X) – MIN(X)---------- (3)

The new values obtained after truncating will be converted into binary form by the following scaling. The values, which are in the range 0 to 5 are converted to 0 and 6 to 10 are converted to1.

Module 2: Training Neural Network:

Following methods will be used to train neural Network for classification purpose.

2.1 Differential Evolution(DE)

DE has a lot of advantages such that it's conceptual simplicity and ease of use. It's specially used to optimize parameters of LS-SVM. It has standard phases like selection, crossover and mutation. DE will be merged with LS-SVM classifier to improve its accuracy.

DE main steps are stated in following Algorithm:

Step 1: Randomly generate a population of N vectors, each of D dimensions.
Step 2: Calculate the objective function value f(Xi) for all target vectors Xi
Step 3: Select 3 points from the population and generate mutant individual Vi using step 1.
Step 4: Apply Crossover operation on each target vector Xi with mutant individual
(generated in step 3) to generate a trail vector Ui using step 2.
Step 5: Calculate the objective function value for vector Ui
Step 6: Choose better of the two (function value at target and trail point) using step 3.
Step 7: Check whether a convergence criterion is met, if yes then stop; otherwise go to step 3

2.2 Least Squares Support Vector Machine (LS-SVM):

LS-SVM classifier is used for finding a hyper plane, which separates various classes. It is improved version of SVM, LS-SVM uses equality constraints, and transforms quadratic programming to linear system solution, which makes the computation and selection of parameters much easier. Therefore, LS-SVM is faster and more robust relative to other regression techniques.

LS-SVM main steps are stated in following algorithm:

Step 1: Load the training data set of n data points, where Xi is the input vector and Yi is the corresponding ith target with values {2,4}.
Step 2: Generate random weights for each input datapoint .
Step 3: Determine the value of the bias term b and initialize the error e for each point randomly.
Step 4: Search for values of e, w and b that minimize the objective function.
Step 5: Construct the Lagrangian function with the solution that must satisfy the Karush- Kuhn-Tucker (KKT) conditions in the set of equations.
Step 6: Calculate number of support vectors
Step 7: Training data for LS-SVM model could be classified with RBD kernel functions.

Module 3: Back Propagation Neural Network:

Neural Networks (NN) are important data mining tool used for classification and clustering. Idea behind BP algorithm is quite simple; output of NN is evaluated against desired output. If results are not satisfactory, connection (weights) between layers are modified and process is repeated again and again until error is small enough. The convergence and network performance depends on many parameters like initial weights, learning rate and momentum used during the training process.

There are three phases in the training process:

1. To send the data using feed forward approach.
2. To calculate the propagated error.
3. To update all weights in the network

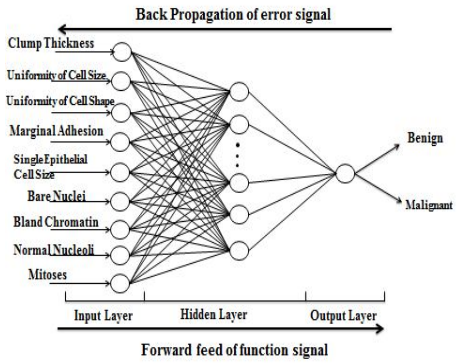

Fig.3 Working of BPN

Feed Forward Computation: During this forward pass the synaptic weights of the networks are all fixed. In this algorithm, actual response will be calculated and compared with desired response. The actual response of the network is subtracted from desired response to produce an error signal.
Error Signal = Calculated Output – Desired Output

Error Back-Propagation: Error Back-propagation algorithm is based on the error correcting learning rule. During this backward pass the synaptic weights all are adjusted in accordance with an error-correction rule. Forward and backward pass are repeated until the error is low enough.
The approximation used for the weight change is given by the delta rule:

$$W_{AB(new)} = W_{AB(old)} - \eta \frac{\partial E^2}{\partial W_{AB}}$$

Where n is the learning rate parameter, which determines the rate of learning, and

$\frac{\partial E^2}{\partial W_{AB}}$ is the sensitivity of the error, *E*2, to the weight *WAB* and determines the direction of search in weight space for the new weight *WAB(new)*.

Module 4: 10-fold cross validation method:

In this study, 10-fold cross validation method will be used for performance evaluation of breast cancer diagnosis using LS-SVM , BPN. k-fold cross validation is a way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Every time, one of the k subsets is used as the test set, and the other k-1 subsets are gathered to form a training set. Then the average error across all k trials is calculated. The advantage of this method is that it is less significant for this method how the data gets divided. Every data point gets to be in a test set only once, and gets to be in a training set k-1 times. As k increases, the variance of the resulting estimate reduces. Steps of 10-fold cross validation performance evaluator :

- Break dataset into 10 sets of size n/10.
- Train on 9 datasets and test on 1.
- Repeat 10 times and take a mean accuracy.

Module 5: Analysis of System:

A confusion matrix contains information about actual and predicted classifications done by a classifier. Performance of system will be evaluated using the data in the confusion matrix.

	Predecited Negative	Predecited Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Fig. 3 Confusion Matrix

The following parameters will be used to analyze the proposed system:

1. **Accuracy–** Percentage of data that are correctly classified to the correct true class

Accuracy=$\sum_{i=1}^{c}$(No. of correctly classified cells in class i) / $\sum_{i=1}^{c}$(Total No. of cells in class i)

2. **Sensitivity –** Percentage of the abnormal data that are correctly classified as abnormal

Sensitivity =TP/(TP + FN)

Where,

TP = True Positive

FN = False Negative (No. of abnormal data classified as normal)

3. **Specificity** - Percentage of the normal data that are correctly classified as normal

Specificity = TN / (TN + FP)

Where,

TN= True Negative(No. of normal data classified as normal)

FP = False Positive (No. of normal data classified as abnormal)

4. **Precision-** It is the proportion of the true positive against all the positive results(both positive tuples and negative tuples)

Precision = TP / (TP + FP)

Where,

TP=True Positive(No. of abnormal data classified as abnormal)

5. **F-score** or **F-measure-** It is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

F-Score = (2 * TP) / (2 * TP + FP + FN)

REFERENCES

[1] Md. Kamrul Hasan, Md. Milon Islam, and M. M. A. Hashem, "Mathematical Model Development to Detect Breast Cancer Using Multigene Genetic Programming", IEEE, 5th International Conference on Informatics, Electronics and Vision , 2016.

[2] Parveen ,Amritpal Singh, "Detection of Brain Tumor in MRI Images, using Combination of Fuzzy C-Means and SVM"IEEE Paper on Signal Processing and Integrated Networks (SPIN) 2015.

[3] Yu-Ling Hou, Chih-Min Lin, Kuo-HsinChen, Te-YuChen ,"Breast Nodules Computer-Aided Diagnostic System Design Using Fuzzy Cerebellar Model Neural Networks", IEEE *Trans.* ON FUZZY SYSTEMS, VOL. 22, NO. 3, JUNE 2014.

[4] Satish Saini ,Ritu Vijay, "Optimization of Artificial Neural Network Breast Cancer Detection System based on Image Registration Techniques," International Journal of Computer Applications,Volume 105 –No. 14, November 2014.

[5] Chandra Prasetyo Utomo, Aan Kardiana, Rika Yuliwulandari , "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques" International Journal of Advanced Research in Artificial Intelligence , Vol.3, No. 7, 2014.

[6] Seema singh ,Sunita Saini, Mandeep Singh, Cancer Detection Using adaptive Neural Network"International Journal of Advancements in Research &Technology,Volume 1, Issue 4, September-2012.

[7] MafazMohsin Al-Anezi , MarwahJasim Mohammed , Dhufr Sami Hammadi ,"Artificial Immunity and Features Reduction for effective Breast Cancer Diagnosis and Prognosis", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, May 2013.

[8] Hassan Jouni1,3, Mariam Issa2, Adnan Harb3 , Gilles Jacquemod1, Yves Leduc1, "Neural Network Architecture for Breast CancerDetection and Classification",IEEE International Multidisciplinary Conference on Engineering Technology (IMCET) ,2016

[9] K.Menaka1, S.Karpagavalli2, "Breast Cancer Classification using SupportVector Machine and Genetic Programming", International Journal of Innovative Research in Computer andCommunication Engineering, Vol. 1, Issue 7, September 2013.

[10] Tanzeem Khan Mansoori,Prof. AmritSuman, Dr. Sadhna K. Mishra"Application of Genetic Algorithm for CancerDiagnosis by Feature Selection", International Journal of Engineering Research & Technology (IJERT) , Vol. 3 Issue 8, August – 2014.

[11] HamidHamidFiuji, Behnaz N. Almasi, Zahra Mehdikhan, Bahram Bibak, Mohammad Pile var, Omid N. Almasi , "Automated Diagnostic System for Breast Cancer Using Least Square Support Vector Machine",American Journal of Biomedical Engineering 2013.

[12] Kalpana Kaushik, Anil Arora, "Breast Cancer Diagnosis using ArtificialNeural Network", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol 7 issue 2 July 2016.

[13] Dr. K.Usha Rani, " Parallel Approach for Diagnosis Of Breast Cancer using Neural Network Technique", International Journal of computer Applications, Vol 10-Nov-3, November 2010.

[14] HamzaTurabieh, "Comparison of NEAT and Backpropagation Neural Network on Breast Cancer Diagnosis",International Journal of computer Applications, Vol 139- Nov-8, April 2016.