International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 10, Number 1 (2017)

Development of Machine Learning Model Using Least Square-Support Vector Machine, Differential Evolutionand Back PropagationNeural Network to Detect Breast Cancer

Madhura D. Vankar
Department of Computer Science AndEngineering,
Shivaji University, D. Y. Patil College of Engineering and Technology,
KasabaBawada,Kolhapur, Maharashtra, India.

Vandana S. Rupnar
Assistant Professor, Department of Computer Science And Engineering,
Shivaji University, D. Y. Patil College of Engineering and Technology,
KasabaBawada , Kolhapur, Maharashtra, India.

Abstract

This paper introduces Development of Machine learning model that classify the input dataset as benign type of breast cancer or Malignant. Data for Machine learning model is retrieved from UCI machine learning repository.The learning model is trained with Breast Cancer Data using feed forward neural network and least square support vector machine.Proposed system includes two phase such as Training phase and Testing phase. In training phase data is preprocessing by using K-nearest neighbor ,feature scaling.K-nearest neighbor is used to replace missing value within data.After completing feature scaling(convert data in the format of zero and one)transfers towards Machine Learning to train feature learning model using LS-SVM , Differential Evolution and BPN classifier.Differential Evolution(DE) is merged with LS-SVM to improve accuracy of LS-SVM classifier and also it requires less training time. It will compare results of two classifiers on the basis of confusion matrix.and provide accurate result to new user in the Testing phase.The effectiveness of Machine learning model is evaluated using 10-fold cross validation method.

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

Introduction

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

Thispaper uses a hybrid classification algorithm usingDifferential Evolution (DE) and Least Squares Support Vector Machine (LS-SVM). LS-SVM technique is used for classification. 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 earlydiagnosis 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.

The Back Propagation algorithm consists of presenting the data, calculating the error, back Propagate the errorand adjusting the weights. The process is repeated multiple times. It is a continuous process of evaluatingoutputs, adapting weights and training with new inputs.

Literature Survey

Fanyu Bu, Yu Ma, Zhikui Chen, Han Xu [1] In paper authors developed a privacy preserving back propagation algorithm depending on the BGV encryption technique on cloud. One property of the designed algorithm is to apply the BGV encryption system to the backpropagation algorithm for preventing disclose of private data with cloud computing. Furthermore, the developed algorithm improved the efficiency of massive data feature learning by incorporating the strong

power of the cloud computing.

Parveen,Amritpal Singh [2]have proposed hybridmethodology 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 systemfor characterizing breast nodules as either benign or malignant on an ultrasonic image.This paper describes a fuzzy cerebellarmodel neural network(FCMNN) as a classifier. It provides high Sensitivity and parameters are decidedhere on the basis of trial and error.

SatishSaini ,Ritu Vijay[4] have describes 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 computed on trial and error method, thereby it's consuming more time.

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

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

Md. KamrulHasan,Md. Milon Islam, and M. M. A. Hashem[7],has developed a 10 fold cross validated mathematical model by using symbolic regression of multigene genetic programming to detect breast cancer.

Problem Definition

To optimize training time and enhance performance Using Machine Learning Model by comparison of Least Square-Support Vector Machine, Differential Evolution and Back Propagation Neural network.

Objectives

The main objectives of this research work are:

- To optimize training time by Merging Differential Evolution with LS-SVM.
- To generate Accurate result by comparison of LS-SVM and BPN using Confusion Matrix.
- It helps to replace missing value in dataset using K-nearest neighbor.
- To evaluate Performance of system with the help of 10-fold cross validation method.

Proposed System Architecture

Our proposed system is especially suitable for big data machine learning with the help training model. Data which is transfer towards feature learning model, processed by using k-nearest neighbor to improve system's performance accuracy.Data converts into binary format by using Feature Scaling.After conversion data is transfer towards learning model to train neural network. It uses two classifier such as LS-SVM and BPN. After comparison of two classifier System, provides accurate result to end-userwith the help of confusion matrix.

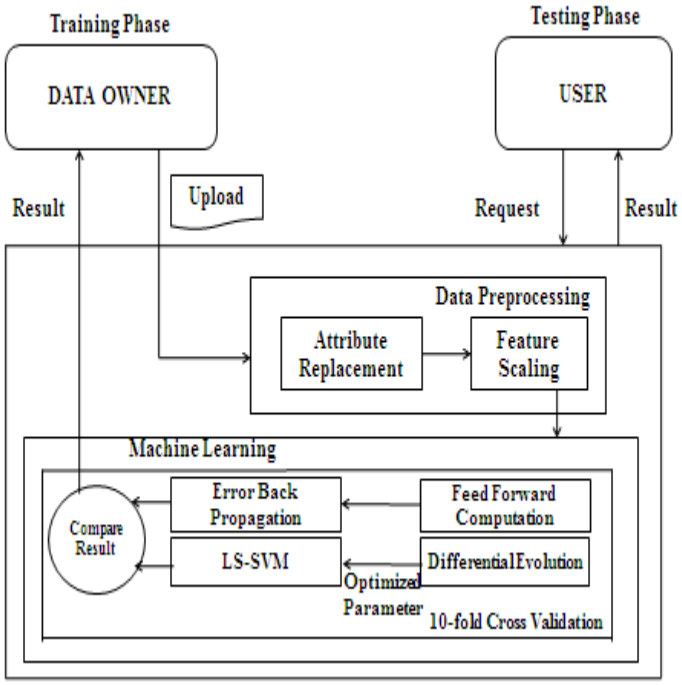

Fig. 1 Proposed System Architecture

Methodology

1. Dataset Used

The proposed algorithm worked 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 that's it is discarded.

	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

Fig.2 Description of Attributes

2. Replacement Of Missing Value Using KNN (K-Nearest Neighbor):

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 is determined using a distance Function.

Distance function can be Euclidean and Manhattan etc. In this work we have considered the Euclidean distance.

K-nearest neighbors algorithm is as follows-

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

$$D(x,y)= \sum_{i=1}^{n} \sqrt{xi^2 - yi2}$$

- After calculating the Euclidean distances choose the data values those having minimum distance. If the value of K is 5 then we have to choose 5 values that having minimum distance.
- Calculate the mean of these chosen values. The mean is given by the equation as:-

  Mean=x1+x2+……xn/n

- Return M as the output value for missing data.

The advantages of KNN imputation are:

K-nearest neighbor can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors).

It can easily treat instances with multiple missing values.

It takes in consideration the correlation structure of the data.

It does not require to create a predictive model for each attribute with missing data. Actually, the k-nearest neighbor algorithm does not create explicit models .

Feature Scaling:

Scaling has the advantage of mapping the desired range of variables rangingbetween minimum and maximum range of network input. Theconversion of the given data sets into binary is done based oncertain ranges, which are defined for each attribute. First from the given range of inputs, theminimum and maximum value is picked up.

This scaling isdone by the following formula:

$$\mathbf{X'} = \frac{X - MIN(X)}{MAX(X) - MIN(X)}$$

The new values obtained after truncating are converted intobinary from by the following scaling.The values, which are inthe range 0 to 5 are converted to 0 and 6 to 10 areconverted to1.

Train Neural Network

Following methods are used to train neural Network.

A]Differential Evolution(DE)

DE has a lot ofadvantages such that it's conceptual simplicity and ease ofuse.It's specially used to optimize parameters of LS-SVM.

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 forall target vectors

Step 3: Select 3 points from the population andgenerate mutant individual using (1)

Step 4: Apply Crossover operation on each targetvector with mutant individual (generated

in step 3) to generate a trail vector using(2)

Step 5: Calculate the objective function value forvector

Step 6: Choose better of the two (function value attarget and trail point) using (3)

Step 7: Check whether a convergence criterion is met,if yes

then stop; otherwise go to step 3

B] Least Squares Support Vector Machine (LS-SVM):

LS-SVM classifiers are one particular sample of SupportVector Machine (SVM). LS-SVM is used for finding ahyper plane, which separates various classes.

LS-SVM main steps are stated in following Algorithm:

Step 1: Load the training data set of n data points,where is the input vector andis the corresponding target with values .

Step 2: Generate random weights for each input datapoint.

Step 3: Determine the value of the bias term b andinitialize the error e for each point randomly.

Step 4: Initialize and using random values.

Step 5: Search for values of e, w and b that minimize theobjective function.

Step 6: Construct the Lagrangian function with thesolution that must satisfy the KKT conditionsinthe set of equations.

Step 7: Calculate number of support vectors

Step 8: Training data for LS-SVM model could beclassified with RBD kernel function

C]Back Propagation Neural Network

Neural Networks are currently research area in medicine. It has a huge application inmany areas such as education, business, medical, engineering and manufacturing. Neural Network plays an important role in a decision support system.

Feed-Forward Computation:

During this forward pass the synaptic weights of the networks are all fixed. In this algorithm, actual response is calculated and compared with desired response. Theactual response of the network is subtracted from desired response to produce an error signal.

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.

K-fold cross validation method

In this study, *10*-fold cross validation method was 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.

Discussion

A confusion matrix contains information about actual and predicted classifications done by a classifier. Performance of such a system is commonly evaluated using the data in the 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) /

$$\sum_{i=1}^{c}(\text{Total No. of cells in class i})$$

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

$$\textbf{Sensitivity} = TP/(TP + FN)$$

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. **Fscore** = (2 * TP) / (2 * TP + FP + FN)

Conclusion

This paperproposed a hybrid classification algorithm to enhance performance of Machinelearning model.It helps to replace missing value from dataset before training.It implements K-nearest neighbor to detect missing value.It helps to optimize training time.It compares the performance between LS-SVM and BPN to predict breast cancer on the basis of confusion matrix. After comparison it provides accurate result to end-user.

References

[1] Fanyu Bu, Yu Ma, Zhikui Chen, Han Xu , "Privacy Preserving Back-Propagation Based on BGV on Cloud" , IEEE 17th International Conference, 2015.

[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," InternationalJournal 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] Md. KamrulHasan,Md. Milon Islam, and M. M. A. Hashem, "Mathematical ModelDevelopment to Detect BreastCancerUsing Multigene Genetic Programming" IEEE Paper2016.

[8]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.

[9]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.

[10]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.

[11]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.

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

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

[14] 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.

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