Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database

Conference Paper · December 2017

DOI: 10.1109/R10-HTC.2017.8289075

CITATIONS
7

READS
428

5 authors, including:

Shaikh Anowarul Fattah
Princeton University
146 PUBLICATIONS **650** CITATIONS

SEE PROFILE

Shajib Ghosh
University of Dhaka
2 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Asir Intisar Khan
Stanford University
28 PUBLICATIONS **174** CITATIONS

SEE PROFILE

Efficient Approaches for Accuracy Improvement of Breast Cancer Classification Using Wisconsin Database

Shajib Ghosh
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
shajibghosh94@gmail.com

Jubaer Hossain
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
jub3110@gmail.com

Dr. Shaikh Anowarul Fattah
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
fattah@eee.buet.ac.bd

Dr. Celia Shahnaz
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
celia@eee.buet.ac.bd

Asir Intisar Khan
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
asir@eee.buet.ac.bd

***Abstract*—Breast cancer is the second leading cause of death for women all over the world. But early detection and prevention can significantly reduce the chances of death. This paper deals with different statistical and deep learning analysis of Wisconsin Breast Cancer Database for improving the accuracy in detection and classification of breast cancer based on different attributes. Applying Naïve Bayes, SVM, Logistic Regression, KNN, Random Forest, MLP and CNN classifiers, higher accuracy is obtained which is up to 98% to 99%.**

Keywords—attributes, algorithms, diagnosis, deep learning, neural network.

I. INTRODUCTION

After increasing at an alarming rate for more than 20 years, breast cancer incidence rates in women began decreasing in 2000, and dropped by about 7% from 2002 to 2003 [1]. But stats have shown that nearly 1.7 million new cases had been diagnosed in 2012 which is second most common cancer overall. This represents about 12% of all new cancer cases and 25% of all cancers in women [2].

Predicting the recurrence of cancer has become a real-world medical problem. Recurrent breast cancer is cancer that comes back in the same or opposite breast or chest wall after a period of time when the cancer could not be detected. Recently, data mining has become a popular and efficient tool for knowledge discovering and extracting hidden patterns from large datasets. It involves the use of sophisticated data manipulation tools to discover previously unknown, valid patterns and relationships in large dataset. Classification algorithms are for finding valuable information in big databases. Data sets having less attributes and higher instances can provide good result [3] than the result we have obtained using the data set from Wisconsin Breast Cancer Database where there are higher attributes and less instances. Too much attributes can misguide a classifier from gaining its maximum result [4,5], which inspired us to use the feature selection method.

Apart from other classifiers, Convolutional Neural Network and Multilayer Perceptron algorithms are also used for better prediction and accuracy. The main objective of our project to find the appropriate classifier model for breast cancer detection and classification with highest accuracy and minimize the cost required for early Mammography Screening.

Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy. Angeline and Dr. Sivaprakasam (2011) [6] compared the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (KNN) to find the best classifier and SVM proves to be the most accurate one with accuracy of 96.99%. In our project, Random Forest has proved to be the best classifier for 50% test data and 50% train data with the accuracy of 96.83%. Guo and Nandi (2006) [7], proposed a Multilayer Perceptron (MLP) as a classifier with retro propagation of error algorithm and obtained an accuracy of 96.21%. Whereas, we have obtained 97.89% accuracy with 5 layers and 10 fold cross validation by using MLP. Karabatak and Cevdet-Ince (2009) [8], presented an automatic diagnosis system for detecting breast cancer based on Association Rules (AR) and Neural Networks (NNs), obtaining a classification accuracy of 97.4%. In our project, we have obtained 98.06% accuracy with 300 feature maps and 10 fold cross validation

978-1-5386-2175-2/17/$31.00 ©2017 IEEE

using Convolutional Neural Network (CNN). Our experiments not only have provided better results than the works mentioned above but also have indicated future scopes to do further research in the field of neural network based classification.

II. Methodology

A. Dataset

The dataset used in this project is obtained from the breast cancer database of the University of Wisconsin Hospitals Madison (Wolberg 1991). There are 11 attributes for each sample. Attributes 2 through 10 have been used to represent instances respectively. Number of instances is 699. But some of the instances are deleted due to missing attributes. There is a class attribute in addition to 9 other attributes. Each instance has one of 2 possibilities: Benign or malignant. One of the other numeric value columns is ID column of instances. Our dataset includes two classes as mentioned earlier. They are benign (B) and malignant (M). We further analyzed data and come up with total 30 attributes with 569 useful data.

The useful attributes are shown in the table below.

TABLE I. Useful Attributes

#	Attributes	#	Attributes
1	The largest or worst of the area - (mean of the three largest values)	16	The largest or worst of the perimeter- (mean of the three largest values)
2	Mean area	17	Mean perimeter
3	Standard error of the area	18	Standard error of the perimeter
4	The largest or worst of the compactness– (mean of the three largest values)	19	The largest or worst of the radius- (mean of the three largest values)
5	Mean compactness	20	Mean radius
6	Standard error of the compactness	21	Standard error of the radius
7	The largest or worst of the concave points (mean of the three largest values)	22	The largest or worst of smoothness - (mean of the three largest values)
8	Mean concave points	23	Mean smoothness
9	Standard error of the concave points	24	Standard error of the smoothness
10	The largest or worst of concavity - (mean of the three largest values)	25	The largest or worst of symmetry - (mean of the three largest values)
11	Mean concavity	26	Mean symmetry
12	Standard error of the concavity	27	Standard error of the symmetry
13	The largest or worst of the fractal dimension - (mean of the three largest values)	28	The largest or worst of texture – (mean of the three largest values)
14	Mean fractal dimension	29	Mean texture
15	Standard error of the fractal dimension	30	Standard error of the texture

B. Classifier Algorithms

The classifier algorithms we have used for data mining are:

1. Naïve Bayes Algorithm,
2. Random Forest Algorithm,
3. K-Nearest Neighborhood Algorithm,
4. Support Vector Machine(SVM) Algorithm,
5. Logistic Regression Algorithm,
6. Decision Tree Algorithm.

Also, we have used artificial neural network techniques **(Multilayer Perceptron–MLP)**, deep learning methods **(Convolutional Neural Network)**, for identifying dominant attributes for detection of benign (B) or malignant (M) classes of breast cancer.

III. Process Description And Observations

A. Features vs Diagnosis Analysis

Different features have different dominance over diagnosis criterion. Here, the figures show the percentage of cancer detection (benign or malignant) according to the feature size or dimensions.

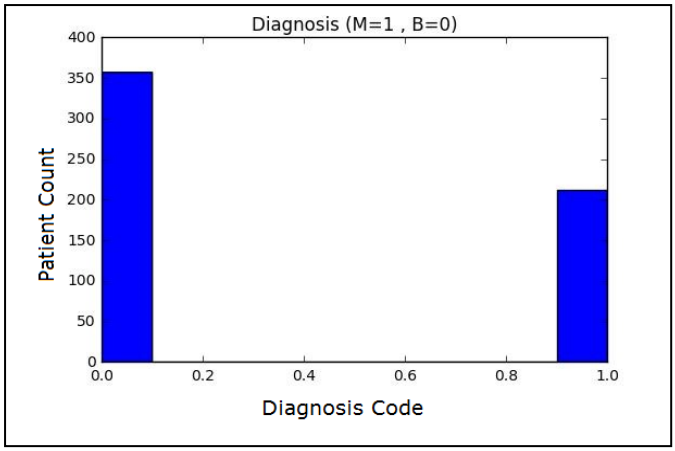

Fig. 1. Diagnosis Code (Malignant = 1, Benign = 0) vs. Patient Count

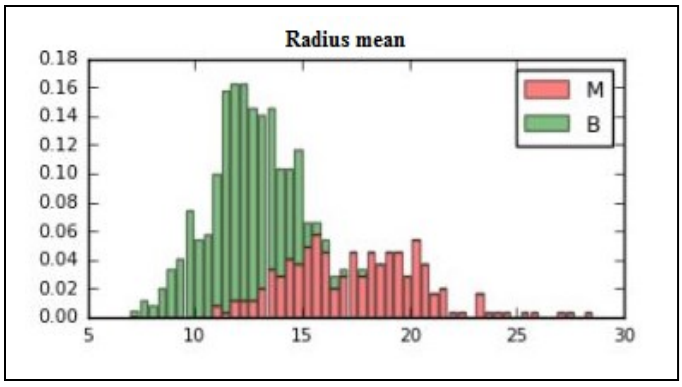

Fig. 2. Mean radius vs diagnosis

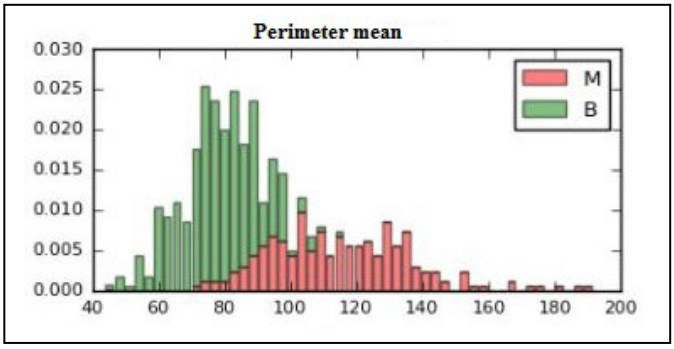

Fig. 3. Mean radius vs diagnosis

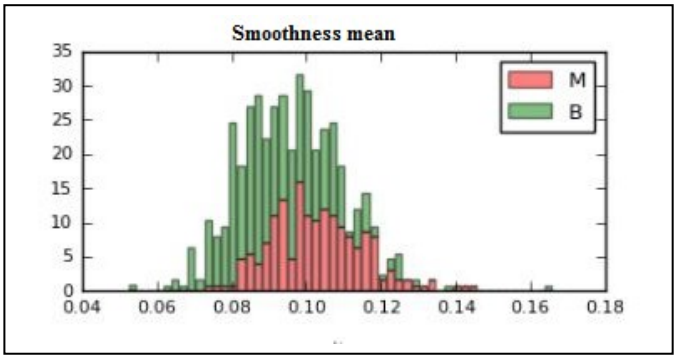

Fig. 4. Mean smoothness vs. diagnosis

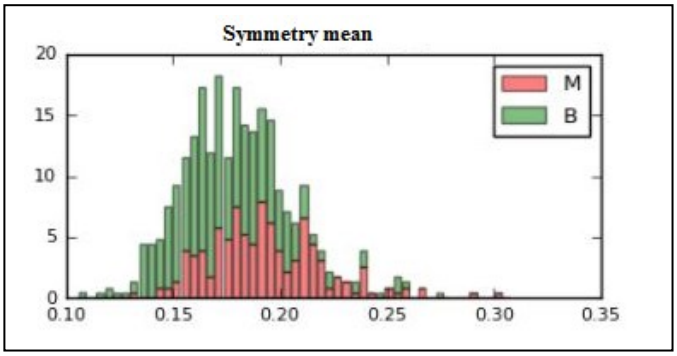

Fig. 5. Mean symmetry vs. diagnosis

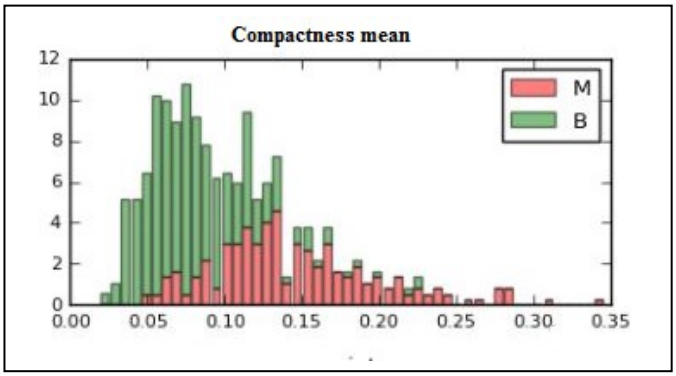

Fig. 6. Mean compactness vs. diagnosis

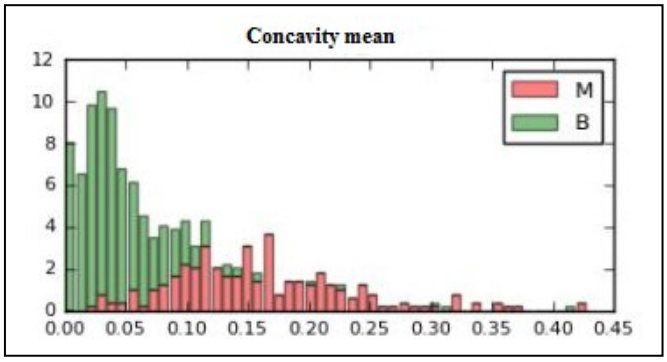

Fig. 7. Mean concavity vs. diagnosis

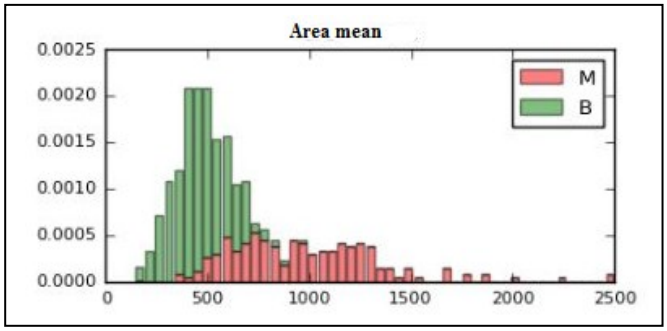

Fig. 8. Mean area vs. diagnosis

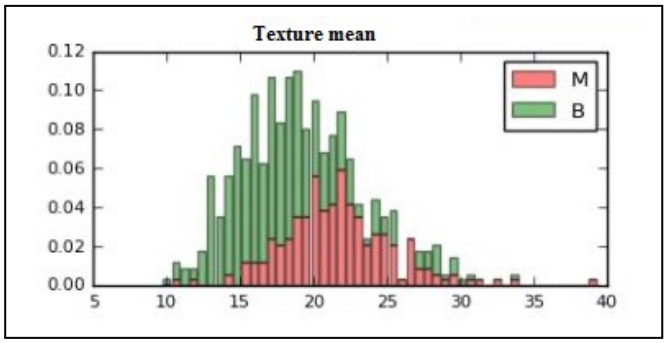

Fig. 9. Mean texture vs. diagnosis

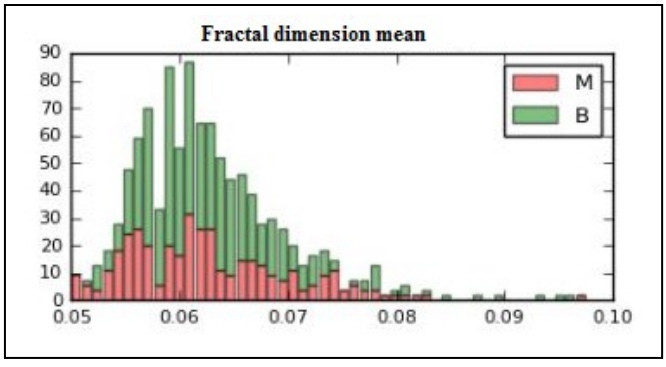

Fig. 10. Fractal dimension mean vs. diagnosis

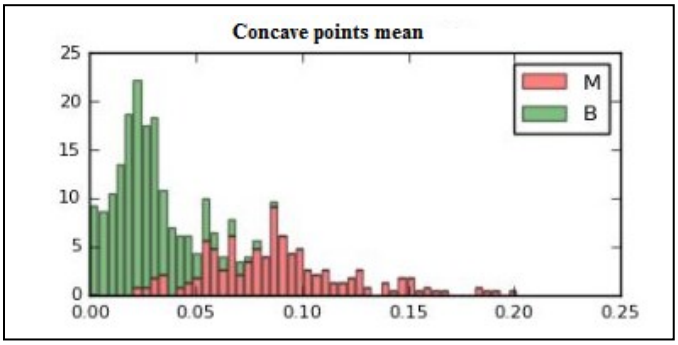

Fig. 11. Mean concave points vs. diagnosis

B. Observations

i) Mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters tend to show a correlation with malignant tumors.

ii) Mean values of texture, smoothness, symmetry or fractal dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further clean up.

The importance of the features in diagnosing whether the tumors are malignant or benign are shown in the following table.

TABLE II. FEATURE IMPORTANCE

Features	Importance (%)
Mean perimeter	20.45614
Mean concave points	19.7067
Mean concavity	18.2823
Mean area	17.3446
Mean radius	11.5543
Mean compactness	4.3632
Mean texture	3.8492
Mean smoothness	1.9023
Mean symmetry	1.6371
Mean fractal dimension	0.9043

C. Comparison Between Different Classifier Models

Different classifier models such as Naïve Bayes, Random Forest, k-Nearest Neighborhood (KNN), Support Vector Machine (SVM) and Logistic Regression are used for binary classification and their performances and accuracy of detection are compared so that the appropriate model can be found.

For this analysis, the total dataset is split into two groups: test data and train data. With the increasing percentage of test data, the accuracy decreases. From the following table and corresponding figure, this result is easily comprehensible.

TABLE III. COMPARATIVE ANALYSIS OF THE PERFORMANCE OF DIFFERENT CLASSIFIER ALGORITHMS

Accuracy (%)						
Data Percentage (%)		Algorithms				
Test Data	Train Data	Naïve Bayes	Random Forest	KNN	SVM	Logistic Regression
10	90	91.2281	100.0	94.7368	98.2456	96.4912
15	85	92.9412	97.6471	96.4706	98.8235	95.2940
20	80	90.3509	93.8596	95.6140	98.2456	95.6140
25	75	91.5493	96.4789	95.7746	97.8873	97.1831
30	70	91.8129	96.4912	95.9064	97.6608	98.8304
35	65	91.4573	96.9849	96.4824	97.4874	91.9598
40	60	90.3509	96.4912	94.2982	96.9298	92.5439
45	55	91.0156	96.8750	92.9688	95.7031	95.3125
50	50	91.9014	96.8310	93.3099	95.7746	95.7746

Graphical representation of the comparative analysis:

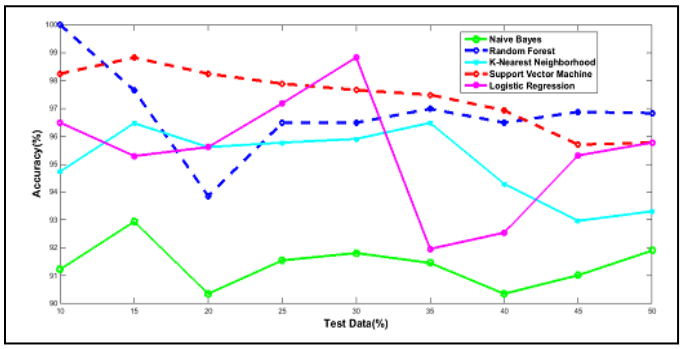

Fig. 12. Graphical representation of the comparative analysis

Above comparative analysis shows that for the highest test data percentage (50%), the highest accuracy is obtained from Random Forest classifier model and the accuracy is 96.8310% which is clearly a significant contribution in early detection of breast cancer.

D. Neural Network and Deep Learning

1. Multi-layer Perceptron (MLP):

A multilayer perceptron (MLP) is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one.

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(X): R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of

features $X = x_1, x_2, ..., x_m$ and a target y, it can learn a non-linear function approximator for either classification or regression [9][10].

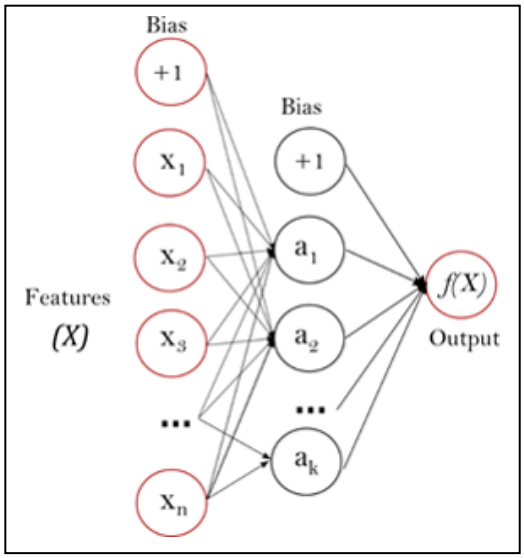

Fig. 13. Multilayer Perceptron (MLP)

We started with one layer, then eventually experimented with more layers. We observed that the convergence time is larger for deeper networks. The result is summarized in the table below.

TABLE IV. SUMMARIZED RESULT OF THE PERFORMANCE OF MULRILAYER PERCEPTRON

Layers	Accuracy (%)	Batch Size	Learning Rate	Learning Momentum	Training Time (epochs)
One	97.5	100	0.03	0.2	100
Five	97.891	100	0.03	0.3	100
Ten	97.5395	100	0.03	0.3	40000

TABLE V. DETAILED ACCURACY OF THE PERFORMANCE OF MULRILAYER PERCEPTRON

Layers	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
One	M	0.948	0.008	0.985	0.948	0.966	0.947	0.989	0.990
	B	0.992	0.052	0.970	0.992	0.981	0.947	0.989	0.990
	Weighted Average	0.975	0.036	0.976	0.975	0.975	0.947	0.989	0.990
Five	M	0.958	0.008	0.985	0.958	0.971	0.955	0.991	0.991
	B	0.992	0.042	0.975	0.992	0.983	0.955	0.991	0.993
	Weighted Average	0.979	0.030	0.979	0.979	0.979	0.955	0.991	0.992
Ten	M	0.948	0.008	0.985	0.948	0.966	0.947	0.991	0.990
	B	0.992	0.052	0.970	0.992	0.981	0.947	0.991	0.993
	Weighted Average	0.975	0.036	0.976	0.975	0.975	0.947	0.991	0.992

All the results are obtained for 10 fold cross validation. And the highest accuracy is 97.891% for five layers which is higher than the accuracy obtained by Guo and Nandi (2006) [7].

2. Convolutional Neural Network (CNN):

Convolutional Neural Networks (CNN) are biologically-inspired variants of MLPs. CNNs use a variation of multilayer perceptrons designed to require minimal pre-processing[11].

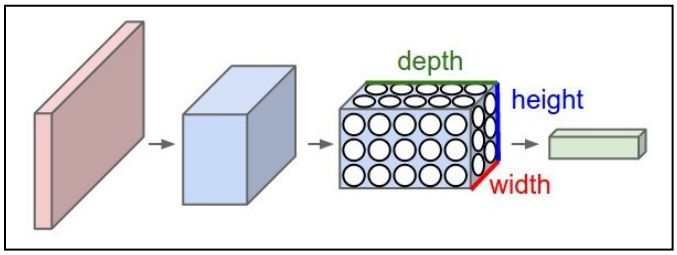

Fig. 14. Convolutional Neural Network (CNN)

For different feature maps and with same batch size and cross validation score we found different accuracies. The result is summarized in the following tables.

TABLE VI. SUMMARIZED RESULT OF THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORK

Feature Maps	Accuracy (%)	Batch Size	Learning Rate	Training Time (epochs)
100	97.71	100	0.03	1000
200	97.53	100	0.03	1000
300	98.06	100	0.03	1000
400	97.36	100	0.03	1000

TABLE VII. DETAILED ACCURACY OF THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORK

Feature Maps	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
100	M	0.953	0.008	0.985	0.953	0.969	0.951	0.994	0.993
	B	0.992	0.047	0.973	0.992	0.982	0.951	0.994	0.996
	Weighted Average	0.977	0.033	0.977	0.977	0.977	0.951	0.994	0.995
200	M	0.953	0.011	0.981	0.953	0.967	0.947	0.995	0.994
	B	0.989	0.047	0.972	0.989	0.981	0.947	0.995	0.996
	Weighted Average	0.975	0.034	0.975	0.975	0.975	0.947	0.995	0.995
300	M	0.962	0.008	0.986	0.962	0.974	0.959	0.994	0.993
	B	0.992	0.038	0.978	0.992	0.985	0.959	0.994	0.995
	Weighted Average	0.981	0.027	0.981	0.981	0.981	0.959	0.994	0.995
400	M	0.948	0.011	0.980	0.948	0.964	0.944	0.944	0.993
	B	0.989	0.052	0.970	0.989	0.979	0.944	0.944	0.995
	Weighted Average	0.974	0.037	0.974	0.974	0.974	0.944	0.944	0.995

All the results are obtained for 10 fold cross validation and the highest accuracy is 98.06% for 300 feature maps with batch size 100.

These machine learning methods are chosen because results obtained from these methods have appeared to be more accurate than traditional classifiers. Moreover, there is future scope of implementing these machine learning techniques for bigger data in faster rate.

Our main focus is to choose most suitable classifier model for obtaining the highest accuracy and to find an improvement of the similar previous works on the same database. We used Weka [12] machine learning tool throughout the experiment and achieved our goal. Moreover, this work has potential to pave the way for future research in the relative field.

IV. Conclusion

We have analyzed our data on the basis of Wisconsin Breast cancer database and we experimented with Naive Bayes, SVM, Logistic Regression, KNN, Random Forest Neural Network and CNN classifiers and obtained highest accuracy. We did a deep investigation in the performance of different deep networks on this dataset. For deep networks, we have found that the convergence time significantly increases and it gets harder to optimize the network. In case of convolutional neural network, we have found the best result with three hundred feature maps. The same result might be obtained with different configuration of the network. Our results of CNN classifier (98.06% accuracy) show comparatively better performance in comparison the work of Karabatak and Cevdet-Ince (2009) [8] where the accuracy was 97.4% using Association Rules(AR) and Neural Network(NN). Such comparative analysis on breast cancer classification would provide further encouragement and insights on the efficient approaches for detection of cancer problems.

References

[1] http://www.cancer.org/cancer/breastcancer/detaildguide/breast-cancer-key-statistics

[2] http://www.wcrf.org/int/cancer-facts-figures/data-specificcancers/breast-cancer-statistics

[3] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[4] Wolberg,William H., and Olvi L. Mangasarian, “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.” Proceedings of the national academy of sciences 87.23(1990): 9193-9196.

[5] Zhang, Jianping, "Selecting typical instances in instancebased learning." Proceedings of the ninth international conference on machine learning. 1992.

[6] Angeline Christobel. Y, Dr. Sivaprakasam (2011), "An Empirical Comparison of Data Mining Classification Methods." International Journal of Computer Information Systems,Vol. 3, No. 2, 2011.

[7] Hong G, Nandi AK (2006), "Breast cancer diagnosis using genetic programming generated feature." Elsevier Pattern recognition 39: 980-987.

[8] Karabatak M, Ince MC (2009), "An expert system for detection of breast cancer based on association rules and neural network." Expert Systems with Applications 36: 3465-3469.

[9] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.

[10] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961.

[11] LeCun, Yann. "LeNet-5, convolutional neural networks". Retrieved 16 November 2013.

[12] Weka 3.8.1, An open source data mining software tool developed at University of Waikato, New Zealand