www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 6 June 2015, Page No. 12853-12858

A Computational Approach for Better Classification of Breast Cancer using GeneticAlgorithm

Prakash Bethapudi [1], E.SreenivasaReddy [2], T.Sitamahalakshmi [3]

[1]Computer Science and Engineering, GITAM Institute of Technology,
GITAM University, VISAKHAPATNAM
prakash.vza@gmail.com

[2]Computer Science and Engineering, College of Engineering
Acharya Nagarjuna University, GUNTUR,
edara_67@yahool.com

[3]Computer Science and Engineering, GITAM Institute of Technology,
GITAM University, VISAKHAPATNAM,
tsm@gitam.edu

Abstract: The proposed work presents a competent diagnosis technique in classifying benign and malignant breast cancer cases using Genetic Algorithm. The breast cancer dataset (Wisconsin Breast Cancer (WBC)) was taken fromUCI Machine Learning Repository, center for machine learning and intelligence systems. Using the proposed Genetic algorithmbased on 3-fold cross validation method, and executing on multiple rules, we obtained preeminent classification accuracy of 97.7% which classified more accurately when compared with the other existing systems. The experimental outcomesillustrate that the categorization using genetic algorithm is loftier to the other classifiers which used WBC dataset. All experiments are carried out on MATLAB.

Keywords: Benign, Breast-Cancer,Diagnostic, Genetic-Algorithm, Malignant, Wisconsin-Breast-Cancer(WBC).

1. Introduction

Breast cancer is one of the most epidemics, frequent and principal cause of mortality among women from most of the countries. Accurate detection of breast cancer can increase the survival rate among them. Effective machine learning approaches have been developed to progress the diagnostic competence for breast cancer. Diverse classification algorithms like decision trees, neural networks, support vector machines, fuzzy sets etc., have been used widely in studying the breast cancer datasets. Also bottomless data mining techniques like feature selection, classification techniques and clustering techniques have been used in studying digital mammograms. The technique used by the radiologists to detect breast cancer is Mammography. However radiologists may not always give accurate results. (Generally classification of benign tumor as malignant tumors and malignant tumors as benign tumor are related with these predictions). A variety of computer aided diagnostic tools have been endorsed to help radiologists. A breast cancer dataset from Wisconsin database, consisting of nine attributes have been used extensively to study various data mining techniques

In this paper, we proposed an algorithm named genetic algorithm which generated best results over other techniques used previously in classifying Benign and malignant cancer cases.Data mining and machine learning depend on classification which is the most essential and important task. 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 sources. Many of them show good classification accuracy. In [1],the performance classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART are compared, to find the best classifier in breast cancer datasets (WBC). The result showed that SVM-RBF kernel is more precise than the former classifiers; its results showed an exactness of 96.84% in WBC. In [2], the decision tree classifier (CART) WBC, achieves accuracy of 94.84% in WBC dataset. When using CART with feature selection (PrincipalComponentsAttributeEval), it scores accuracy of 96.99 in WBC dataset. When CART is used with feature selection (Chi-SquaredAttributeEval), it gave an accuracy of 94.56 in WBC dataset. In [3], C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) when competed to find the finest classifier in WBC, SVM proves to be the more perfect with an accuracy of 96.99%.Kamadi V S R P Varma et.al,[4] proposed Genetic Algorithm approach for the early diagnosis of diabetics and achieved an accuracy of 74.9% for initial population of 100 rules. In The proposed system that is, by using Genetic Algorithm on WBC dataset with basic 11 features, based on 3-fold cross validation method, and implementing on various rules, we obtained perceptible classification accuracy of 97.7% which classified more accurately when compared with the other existing systems. The results of current technique and the existing techniques in classification of benign and malignant cases are shown with their accuracies in percentages. The rest of the paper is organized as follows; the Section II describes the dataset description. The proposed model and flow chart is presented in Section III. Results were discussed in Section IV and concluded with Section V.

2. Dataset Description

For this work, we considered the original Wisconsin Breast Cancer datasets. These were taken from the UCI Machine Learning Repository [5], to distinguish malignant (cancerous) from benign (non-cancerous) cases. The total dataset consists of 699 instances of which each instance contains 11 features / attributes in total, including the id number and class fields. A brief description of dataset and the attributes existing in this datasets is presented in table1 and table2.

Table1 Breast Cancer Dataset Description

Dataset	**No. of Attributes**	**No. of Instances**	**No.of Classes**
Wisconsin Breast Cancer (Original)	11	699	2

Table2Wisconsin Breast Cancer Dataset Attributes

S.No	**Attribute**	**Domain**
1	Sample code number	id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 - 10
11	Class	2 - benign, 4 - malignant

3. Proposed Model Description

In the proposed model, we collected the original dataset of Wisconsin breast cancer dataset. This data has been preprocessed. Any gaps in the fields may be filled with the mean value of that particular field. As already discussed this dataset consists of eleven attributes in total. First we have classified all the 699 instances into two classes. The first class is classified as cancerous and the second class is classified as non-cancerous cases. From these two classes we extracted three folds using some technique. We may extend the folds according to our wish as folds3, folds5, folds10 and so on. Now from these folds we prepare a training dataset as well as testing dataset. We provide the train dataset and the test dataset to the proposed genetic algorithm and apply rules on them.

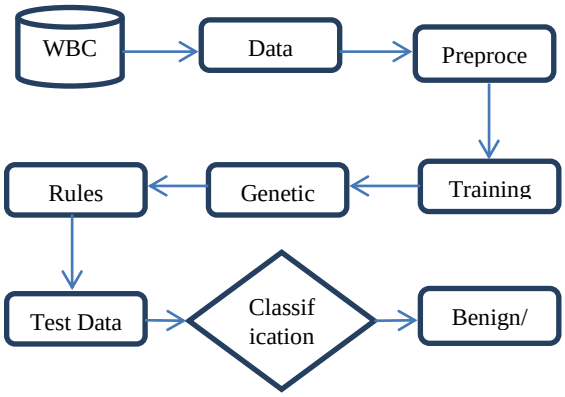

Figure 1. Block Diagram of the Proposed Model

Finally by using test data and classification technique we got the accuracy of 97.7% on the data which is more promising then the existed results from various techniques. Confusion matrix has been constructed for the three extracted folds. The accuracy, sensitivity and specificity for rules on each fold are retrieved. This is carried out for multiple times using different ranges of rules. Here we considered the folds for 50, 100, 150 and 200 rules for all the three folds using the genetic algorithm. We examined exceptional results when compared to the existing results which have been retrieved by various techniques. The accuracy, sensitivity and specificity of the proposed model are presented in the table 3, 4 and 5. And the performance of the proposed model for folds using confusion matrix is shown in below Figures 2, 3 and 4. With the proposed model we achieved an average accuracy 97.7% which is presented in the table 5.

3.1 Genetic Algorithm

1. [Start] Generate population of n chromosomes randomly
2. [Fitness] calculate fitness of every chromosome in the population
3. [New population] Repeat below steps and Construct new population until the new population is complete
    i. [Selection] Based on their fitness, select two parent chromosomes from the population (the better fitness, the bigger chance to be selected)
    ii. [Crossover] Apply crossover on the parents to form a new offspring (children). If no crossover, then offspring will be an exact copy of parents.
    iii. [Mutation] Apply mutation on new offspring at each locus.
    iv. [Accepting] dwell new offspring in a new population
4. [Replace] Use newly generated population for further run of algorithm

5. [Test] If the end condition is fulfilled, terminate and return the best solution in current population

6. [Loop] Else go to step 2

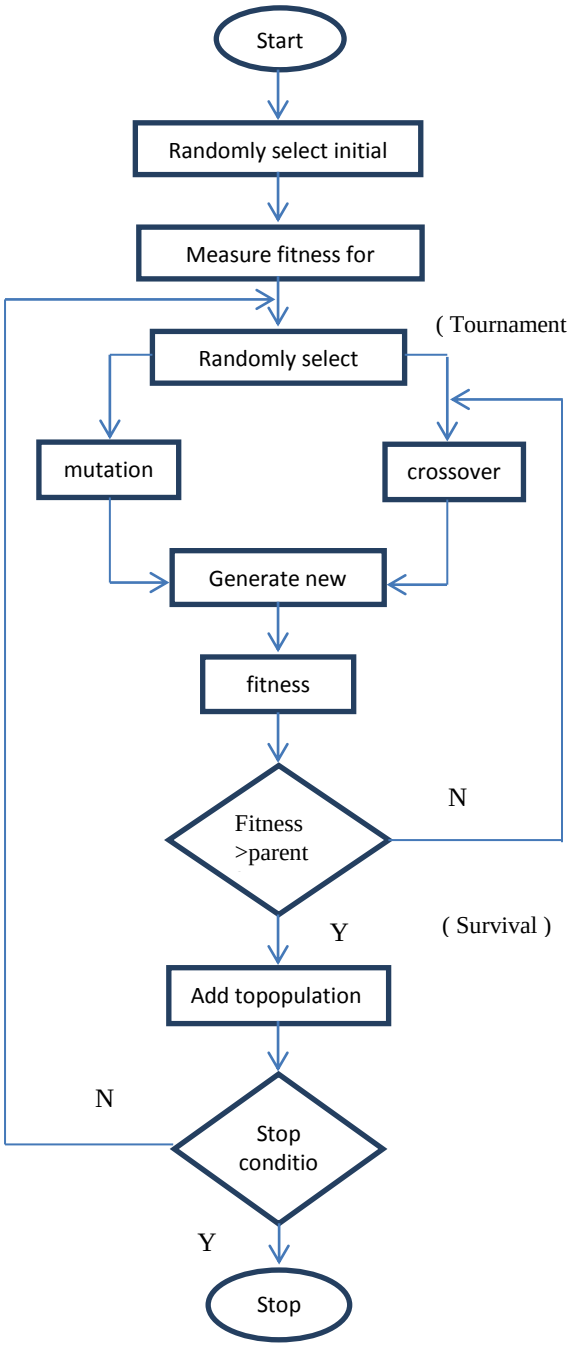

Figure 2. Block Diagram of Genetic Algorithm

3.2 Operators of GA

The most important operators of the genetic algorithm are crossover and mutation. The performance is controlled mainly by these two operators.

3.3 Encoding of a Chromosome

The chromosome should contain resultant information which it represents. Commonly used way of encoding is a binary string. The chromosome looks as follows:

Chromosome 1	1100110 0100110110
Chromosome 2	11011110 0 0 011110

Each chromosome contains one binary string and each bit in this string represents some characteristic of the result. There are many other ways of encoding whichmainly depends on the problem.

3.4 Crossover

We use crossover operator to combine two parent chromosomes and to produce new offspring chromosomes. The main idea behind this crossover is that the new chromosomes may be better than both parents when best characteristics are taken from each of the parent. Crossover operators can be done in many ways like One-point crossover, two point crossover, Uniform crossover, Arithmetic crossover andHeuristic crossover. After deciding encoding we consider crossover. Crossover selects genes from parent chromosomes and creates a new offspring.

3.4.1.1 One-point crossover

The simplest way to perform one point crossover is choose randomly some crossover point and copy everything before this point from a first parent and then copy everything after a crossover pointfrom the second parent. The crossover looks as follows

Chromosome 1	11011 | 00100110110
Chromosome 2	11011 | 11000011110

After interchanging parents based on crossover points, the obtained offspring's with crossover point '|' looks as

Offspring 1	11011 | 11000011110
Offspring 2	11011 | 00100110110

3.4.1.2 Two-point crossover

This operator selects two crossover points randomly within the chromosome and then the parents are interchanged between these points to produce new offspring's.Consider crossover for two chromosomes.

Chromosome 1	11011 | 0010011 | 0110
Chromosome 2	11011 | 1100001 | 1110

By using crossover point, offspring's are produced by interchanging the parent chromosomes.

Offspring 1	11011 | 0010011 | 0110
Offspring 2	11011 | 0010011 | 0110

3.4.1.3 Uniform Crossover

This operator decides which parent will contribute in offspring chromosomes. The crossover operator allows parent chromosomes to mix at gene level rather than segment level. For this fetch two parents for crossover.

Chromosome 1	1101100100110110
Chromosome 2	1101111000011110

If mixing ratio is 0.5, then half of the genes in offspring come from part1 and other half comes from part2. Offspring after uniform crossover would be

Offspring 1 $1_1 1_2 0_2 \;\; 1_1 1_1 1_2 1_2 0_2 0_1 0_1 0_2 1_1 \;\; 1_2 1_1 1_1 0_2$

Offspring 2 $1_2 1_1 0_1 \;\; 1_2 1_2 0_1 0_1 1_1 0_2 0_2 1_1 1_2 \;\; 0_1 1_2 1_2 0_1$

The subscript in the above offspring notation indicates from which part the gene came.

The sequential steps for crossover operation:

1. Select two parent rules from the tournament selection process.
2. Select a random point in the individual parent expression.
3. Exchange the sub parts front and rare parts at the selected point
4. Find the fitness of the newly formed rules; if the fitness of the off springs is maximum value add these rules to the initial rules.
5. Repeat the above process for required number of times

There are other ways to make crossover by choosing more crossover points. Crossover may be complicated and depends on encoding of chromosomes. Specific crossover made for a specific problem can improve performance of the genetic algorithm.

3.5 Mutation

After a crossover is performed, mutation takes place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1. The mutation depends on the encoding as well as the crossover. For example when we are encoding permutations, mutation could be exchanging two genes

Steps involved inmutation operation with attribute modification:

1. Select a random point within the attribute range.
2. Form the new rule by changing the selected attribute value.
3. Compute the fitness of the newly formed rule if the fitness is greater than the parent then add this rule to the initial population
4. Repeat the above process for required number of times

Mutation forms:

Original Offspring 1 1101111000011110
Original Offspring 2 1101100100110110

Mutated Offspring 1 1100111000011110
Mutated Offspring 2 1101101100110110

The accuracy is calculated on fold1 using a specified set of rules on it and the percentage of accuracy have been retrieved. Similarly the sensitivity and specificity too are calculated on fold1 using the set of specified rules say 50 for first run and the percentage of sensitivity and specificity have been calculated. In the same way Accuracy, sensitivity and specificity have been calculated by applying 100, 150 and 200 rules on fold1 and the percentage of accuracy were calculated for fold1. The same process is carried out for fold2 and fold3 by applying 50, 100, 150 and 200 rules and the percentage of accuracy, sensitivity and specificity have been calculated. The obtained results were compared with the results of previous methods applied on Wisconsin original breast cancer dataset. The results obtained by the proposed system proved to be more encouraging when compared to the results of the previous systems. Confusion matrix consists of actual values as confirmed by the experiment and predicted values which are predicted by the test. This consists of positive and negative values say true positive and false positive foe positive values and false negative and true negative for negative values. From these values the sensitivity, specificity and accuracy for the values are calculated and plotted in confusion matrix.Table 3 presents formulas for Accuracy, Sensitivity and Specificity measures.

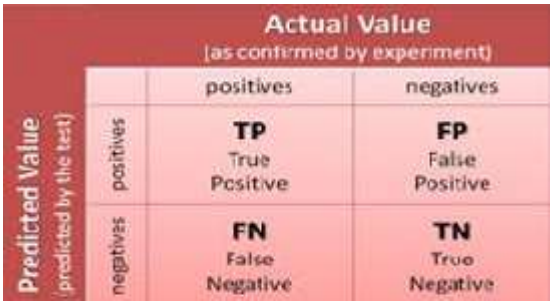

Figure3. Confusion Matrix

Table3 Performance Metrics

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$

4. Results and Discussion:

The proposed Genetic Algorithm used Wisconsin Breast cancer dataset from UCI machine learning repository. We used 3-folds for the process. From the folds we retrieved three train datasets and three test datasets. Each train data is considered from two folds which contain 466 records and each test data is considered from one fold which contains 233 records. Executing the train and test data by using Genetic Algorithm on multiple rules, we obtained preeminent classification accuracy of 97.7% which gave more accurateresults when compared with the other existing systems. The results are presented in Figures 4, 5 and 6, and the classificationAccuracy, Specificity and Sensitivity are presented in Table 4. The Accuracy of each fold and average accuracy is presented in Figure 7. And the comparisons of existing and proposed system arepresented in Table 5 and Figure 8.

Figure 4. FOLD-1

Figure 5. FOLD-2

Figure 6. FOLD-3

Table 4Average Accuracy, Specificity andSensitivityresults obtained with three Folds

Initial Rules – 200			
	Accuracy	**Specificity**	**Sensitivity**
Fold-I	98.7%	100%	98.0%
Fold-II	96.6%	100%	95.0%
Fold-III	97.9%	100%	96.8%
Average	**97.7%**	100%	96.6%

The overall Graphical representation of Accuracy, for three folds independently for 200 initial rules and the average accuracy of all the folds is shown in the below graph.

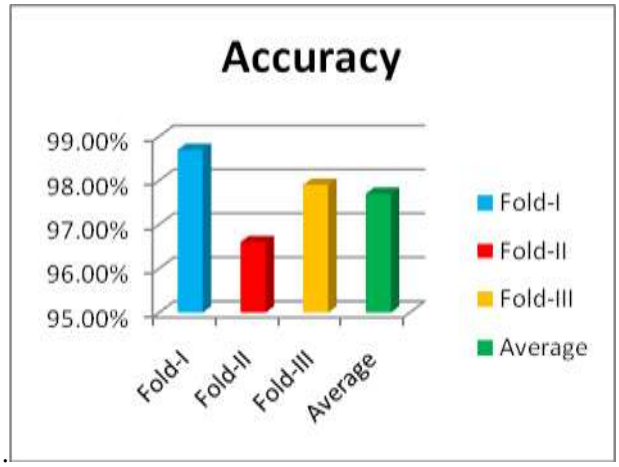

Figure7. Accuracy measure obtained in 3 folds

Table 5 Comparisons Of Existing And proposed Experimental Results

MethodReference	**Classifier**	**Classification accuracy**
S.Aruna et.al[1]	SVM-RBF kernel	96.84%
D.Lavanya et.al [2]	SVM	96.99%
Angeline Christobel et.al [3]	CART with feature selection (Chi-square)	94.56%
Gouda I. Salama et.al [6]	SMO+J48+NB+IBk	97.2818%
Ming-Feng Han et.al [7]	CNFS	97.4%
SoumadipGhosh et.al [8]	MLP BPN	95.71%
Charoenchai Sirisomboonrat et.al [9]	C4.5	94.72%
ProposedModel	**Genetic Algorithm**	**97.7%**

The experimental results obtained in various classifiers are presented in Table 5. The comparison classification accuracy of different methods is presented in Figure 8. It seemed that the proposed Genetic Algorithm proved to be more accurate in classification when compared with other existing models.

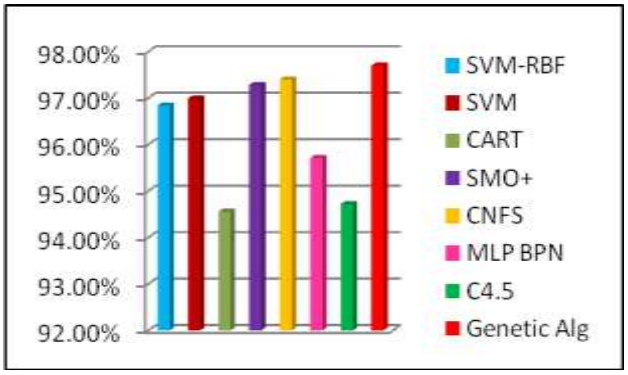

Figure 8. Classification comparison between existing methods and proposed method

5. Conclusion

From the results obtained for various rules when applied, it is very clear that the proposed Genetic Algorithm is able to classify the Type 1(Benign) cancer cases with an accuracy of 100% whereas the type 2(Malignant) cases are classified up to 96.6% accuracy. Result shown in figures 1,2 and 3. Hence by using the proposed method, early detection of cancer cases are possible and detected accurately which helps the patent to take better treatment and the radiologist to provide effective treatment for the patents. Also the average classification accuracy obtained is 97.7 which are more accurate than the existing methods shown in table5.The proposed system used three folds but can use more number of folds say 5, 10, 15… for more encouraging results which may be carried out as our further work. Our further work also includes in applying the same technique on other breast cancer data sets and find out the classification accuracy.

Acknowledgments

My thanks to UCI Machine Learning Repository for providing the datasets of Wisconsin breast cancer patients. My sincere thanks to GITAM UNIVERSITY, VISAKHAPATNAM, INDIA, and our Head of the department Prof P V Nageswar Rao for providing me the necessary software and resources in carrying out the research work. Last but not the least my heartfelt thanks to my dear kids Meghana and Yashwant and my wife Jaya Kalyani for helping me in proceeding with my work smoothly.

References

1. S.Aruna et.al.(2011) Knowledge based analysis of various statistical tools in detecting breast cancer.
2. D.Lavanya, Dr.K.Usha Rani,..," Analysis of feature selection with classification: Breast cancer datasets",Indian Journal of Computer Science and Engineering (IJCSE),October 2011
3. 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.
4. Kamadi VSRP Varma, Dr. AllamApparao, Dr. T. SitaMahalakshmi, Dr. P.V .NageswarRao, K Narasimharao, NCETCS'14 (pages289-295) A Computational Intelligence technique for effective diagnosis of diabetes disease using Genetic Algorithm
5. Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
6. Gouda I. Salama, M.B.Abdelhalim, and MagdyAbd-elghanyZeidijcit(2277-0764) vol.1, issue.1 ,2012. Breast Cancer Diagnosis on Three Different Datasete using Multi-Classifiers
7. Ming-Feng Han, Chin-Teng Lin, Fellow, IEEE, Jyh-Yeong Chang. (2010.IEEE). A Compensatory NeuroFuzzy System with. Online Constructing and Parameter Learning
8. SoumadipGhosh, Member,IEEE, SujoyMondal, Member,IEEE, BhaskarGhosh , Student Member. A Comparative Study of Breast Cancer Detection based on SVM and MLP BPN Classifier
9. CharoenchaiSirisomboonrat, Krung Sinapiromsaran. Breast Cancer Diagnosis using Multi-Attributed Lens Recursive Partitioning Algorithm.(2012 International conference ICT and Knowledge Engineering)