ENHANCED BREAST CANCER RECOGNITION BASED ON ROTATION FOREST FEATURE SELECTION ALGORITHM

[1]Indrajit Mandal**, [2]Sairam.N

[1,2]School of Computing
SASTRA University, India-613401
[1]indrajit@cse.sastra.edu**, [2]sairam@cse.sastra.edu

Abstract

Optimization problems are dominantly being solved using Computational Intelligence. One of the issues that can be addressed in this context is problems related to attribute subset selection evaluation. This paper presents a computational intelligence technique for solving the optimization problem using a proposed model called Modified Genetic Search Algorithms (MGSA) that avoids local bad search space with merit and scaled fitness variables, detecting and deleting bad candidate chromosomes, thereby reducing the number of individual chromosomes from search space and subsequent iterations in next generations. This paper aims to show that Rotation forest ensembles are useful in the feature selection method. The base classifier is multinomial logistic regression method integrated with Haar wavelets as projection filter and reproducing the ranks of each features with 10 fold cross validation method. It also discusses the main findings and concludes with promising result of the proposed model. It explores the combination of MGSA for optimization with Naïve Bayes classification. The result obtained using proposed model MGSA is validated mathematically using Principal Component Analysis. The goal is to improve the accuracy and quality of diagnosis of Breast cancer disease with robust machine learning algorithms. As compared to other works in literature survey, experimental results achieved in this paper show better results with statistical inference.

Keywords:

Computational Intelligence, attribute subset selection, Rotation forest, Haar wavelets, Modified Genetic Search algorithm.

1. Introduction

Computational Intelligence is a branch emerged from artificial intelligence. The Genetic Algorithm (GA) is a heuristic method of evolutionary algorithm based on the idea of survival of the fittest individual [7, 12,18]. GA is used as optimization algorithms which imitates natural

Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 223–234, 2012.

DOI : 10.5121/csit.2012.2322

evolution process. It is a stochastic process method that works on the natural selection strategy and natural genetics. It can be applied to a variety of wide range of problems. The basic steps involved in GA are detailed in [2, 3, 7, 9, 11,16]. The GA has gained popularity due to the inherent advantages [15] that it provides for the computation such as it does not have much mathematical requirements [17,19,20], evolution operators performing global search and its adaptability in hybridizing with domain dependent heuristics.

2. Modified Genetic Search Algorithm

As GA finds solution through evolution process [18], it is not inclined to good solution but move away from bad solutions. There is a chance that it may lead solution to a dead end. Also the number of iterations or generations required is very high. In the proposed MGSA, the idea is to identify and eliminate the bad chromosomes using merit and scaled variables so that the search space is reduced. Once the search space is minimized, containing prospective chromosomes then it leads to better optimization in search process. Here the merit and scaled variables refer to the classification error rates of chromosomes. Bad chromosomes indicate those individuals which lead to a dead end. Here is a brief summary of MSGA:

1) **Initial population generation.** The parameters of the model to be optimized are considered as chromosomes. A randomly chosen set of population of individuals is a binary string with a fixed length.

2) **Fitness function:** In each generation for which GA is run, the fitness of each individual is determined which is close to optimality. Here in the attribute selection problem, we have defined linear function

f `= af + b**, where **f `, f are scaled and raw fitness values of chromosomes and **a,b** are constants.

3) **Detection and Elimination:** Those chromosomes having greater classification error rate determined by merit and scaled variables are determined and eliminated so that it is not allowed to be a candidate chromosome in the next generation and thereby filtering out the bad chromosomes and hence the number of iterations in the subsequent generations gets reduced.

4) **Selection:** A pair of best fit chromosomes or individuals which has least classification error rate is selected from the residual population.

5) **Crossover:** It is the reproductive stage where two new individuals are crossed with probability Pc to generate new pair of offspring.

6) **Mutation:** A single point alteration of bit string from zero to one and vice-versa with probability Pm is applied to selected individuals to generate a new pair of offspring and avoids premature convergence.

3. Experimental design and results

The experimental setup exploits the use of Modified Genetic Search algorithm to optimize a subset of inputs used to classify patients as having either benign or malignant forms of breast cancer tumors using java programming. Benign tumors are non-progressive and harmless whereas malignant tumors spread rapidly and very harmful. The real time data is adapted from Breast Cancer database [1] that contains 683 instances after deleting 16 records containing one or more missing values. Besides it contains nine numeric inputs and a target attribute class which takes on values 2(benign) and 4(malignant). Ten attributes are Clump_Thickness, Cell_Size_Uniformity, Cell_Shape_Uniformity, Marginal_Adhesion, Single_Epi_Cell_Size, Bare_Nuclei, Bland_Chromatin, Normal_Nucleoli, Mitoses, Class. These attributes are used in pathology report on fine needle aspirations to determine whether a lump in a breast could be either malignant (cancerous) or benign (non-cancerous). The details of these parameters are discussed in [15]. The data distribution for Class and indicates that 65% (445/683) of records have value 2(benign) while the remaining 35% (238/683) have value 4(malignant). The detail of the feature ranking is shown below:

3.1. Feature selection

Rotation forest is an ensembles classifier based on feature extraction. The heuristic component is the feature extraction to subset of features and rebuilding a total feature set for each classifier. We have used ensemble that consists of multinomial logistic regression model with a ridge estimator as classifier [21] and Haar wavelets as projection filter. To our knowledge from literature, this ensemble is used for first time. We have found experimentally that our ensemble gives better result compared to the ensemble discussed by Juan [22].

The proposed ensemble is detailed below:-

Let $\mathbf{x=[x_1, \ldots ,x_n]^T}$ be an instance given by **n** variables and **X** be the training sample in a form of **N x n** matrix. Let vector $\mathbf{Y=[y_1,\ldots.y_N]}$ be class labels , where $\mathbf{y_j}$ takes a value from the set. Let $\mathbf{D_1,\ldots..,D_L}$ be the classifiers in ensemble and **F** is feature set.

In ensemble learning, choosing **L** in advance and training classifiers in parallel is necessary.
Follow the steps to prepare the training sample for classifier **Di**:

1. Split **F** randomly into **K** disjoint or intersecting subsets. To maximize degree of diversity, disjoint subsets are chosen.

2. Let $\mathbf{F_{i,j}}$ be jth subset of features to train set of classifier Di.
Draw a bootstrap sample of objects of size 75 percent by selecting randomly subset of classes for every such subset. Run Haar wavelet for only **M** features in $\mathbf{F}_{i,j}$ and the selected subset of X. Store the coefficients of the Haar wavelets components, $\mathbf{a_{ij[1]},\ldots..a_{i,j[Mj]}}$, each of size **M x 1.**

3. Arrange obtained vectors using coefficients in a sparse “rotation” matrix **Ri** having dimensionality **n x ∑Mj**. Compute the training sample for classifier Di by rearranging the columns of **Ri**. Represent the rearranged rotation matrix $\mathbf{R^a_i}$ (size N x n). So the training sample for classifier **Di** is $\mathbf{X\ R^a_i}$.

Table 1 Attribute ranking by Rotation forest ensemble comprising multinomial logistic regression model with a ridge estimator with Haar wavelets as projection filter using 10 fold cross validation. The ranks are generated using Ranker search method.

average merit	**average rank**	**attribute**
92.704 +- 0.479	**1.3 +- 0.46**	**Cell_Size_Uniformity**
92.275 +- 0.455	**1.7 +- 0.46**	**Cell_Shape_Uniformity**
90.351 +- 0.524	**3 +- 0**	**Bland_Chromatin**
88.968 +- 0.697	**4 +- 0**	**Bare_Nuclei**
87.554 +- 0.251	**5.1 +- 0.3**	**Single_Epi_Cell_Size**
86.663 +- 0.324	**6.6 +- 0.66**	**Normal_Nucleoli**
86.409 +- 0.442	**7 +- 0.89**	**Marginal_Adhesion**
86.123 +- 0.444	**7.3 +- 0.9**	**Clump_Thickness**
78.97 +- 0.321	**9 +- 0**	**Mitoses**

3.2 Case 1:

The Naïve Bayes [5] is applied to the dataset to classify with 10-fold cross validation [10] and shows that it achieves a very impressive 96.3397 %(658/683) classification accuracy. Details are shown in table2.

A confusion matrix [4, 10] shown in table 2 contains the analytical details of classifications where nine attributes are input to it. Its performance is evaluated based on the data in the matrix for two class classifier. Accuracy is measured by Received Operator Characteristics (ROC) [4] area under graph with TP as Y-axis and FP as X-axis and ranges from zero to one. With area=1 represents perfect test. For class benign ROC plot =0.99 is shown in fig1 for case 1 and Cost/benefit analysis is shown in fig 2 where gain is 0.7. When relatively the number of negative instances are higher than positive instances, then F-measure gives better understanding of accuracy level and is computed as $F = [(\beta^2+1) * P * TP] / [\beta^2*P+TP]$ where β varies from zero to infinity and is used to balance weight assigned to TP and P. Higher the value of F-measure higher is the accuracy level of classifier and its value ranges from zero to one. The following tables 2,3 show the confusion matrix and accuracy parameters respectively.

3.3 Case 2:

In case 1 all the nine attributes are considered but in real world data irrelevant, redundant or noisy attributes are common phenomena, which impairs the result. The learning scheme Wrapper subset evaluation [8] with Naïve Bayes classification [8] is now integrated with Genetic Search algorithm [5]. After filtration process only seven attributes are selected as relevant. Attributes single_cell_size and mitosis is eliminated. The attributes of Genetic algorithm [23] that includes a population size of n=20 chromosomes, crossover probability P_c=0.6 and mutation probability P_m=0.033.As specified, Genetic Search algorithm creates an initial set of 20 chromosomes. Now reclassifying the records using naïve Bayes with 10-fold cross validation [13]; however, this time only seven attributes are input to the classifier. Selected attributes: 1,2,3,4,6,7,8. They are Clump_Thickness, Cell_Size_Uniformity Cell_Shape_Uniformity, Marginal_Adhesion, Bare_Nuclei, Bland_Chromatin, Normal_Nucleoli.

In table 6, every subset is a chromosome and merit is the fitness score reported by naïve Bayes, which is equal to the corresponding classification error rate. Also, each chromosome's scaled fitness is shown in the scaled column where we use linear scaling technique to scale the values. By definition, the raw fitness and scaled fitness values have the linear relationship

$$\mathbf{f` = af + b} \qquad \mathbf{(1)}$$

where f ` and f are the scaled and raw fitness values respectively. The constants **a** and **b** are chosen where

$$\mathbf{\hat{f}_{avg} = f_{avg} \text{ and } \hat{f}_{max} = K\,\hat{f}_{avg.}} \qquad \mathbf{(2)}$$

The constant K represents the expected number of copies of the fittest individual in the population. Thus, by computing the average fitness values from table 6, we obtain $\hat{f}_{avg}$ =0.055755 and f_{avg} =0.055753. To find a, b the fitness values from last two rows in table 6 are chosen to solve simultaneous equations:

$$\mathbf{0.05911= 0.05417a + b} \qquad \mathbf{(3)}$$

$$\mathbf{0.06677 =0.04392a + b} \qquad \mathbf{(4)}$$

Solving equations (3),(4) we get a= -0.747317and b= 0.0999592. We use equation (2) to determine K i.e. K= $\hat{f}_{max}$ / $\hat{f}_{avg}$ =0.07006/0.056999= 1.2291.

Observe in fifth row in table 6 $\hat{f}$=0. The raw fitness value of 0.13324 corresponds to the largest classification error in the population produced by chromosome {8}, and as a result, $\hat{f}$ is mapped to zero to avoid the possibility of producing negatively scaled fitness. Here the chromosome {8} is detected and deleted so that it does not propagate to the next generation.

The improved classifier has accuracy level of 96.9253 %, which indicates that the second model outperforms the first model by 0.5856 % where the input to the later model is only seven attributes indicated in ROC area. For class benign ROC plot =0.993 is shown in fig3 and Cost/benefit analysis is shown in fig 4 where gain is 0.7 and observe the smoothness achieved in the graph. Moreover reducing the number of attributes in second model did not affect the gain in cost benefit analysis. The plot matrix of selected attributes is shown in fig5. That is, the classification accuracy has improved positively where only seven of nine attributes are specified in the input. Though the increased accuracy is not a dramatic improvement, it shows the strength of Modified Genetic Search algorithm is an appropriate algorithm for attribute selection process.

3.4 Case 3:

Now to validate the result obtained in case 2, Principal Component Analysis (PCA) [6] with attribute selection [14] is applied to the data set consisting of nine attributes for the attribute selection process. It can be noticed that the accuracy of Modified Genetic Search algorithm integrated with Naïve Bayes is mathematically proven by comparing the results with the output from Principal component analysis like the former result only seven attributes are selected whereas in PCA analysis result. Correlation matrix [14] of the entire data set is depicted in table 7 which gives the correlation between all pairs of attributes. Eigen vectors shown in table 8 gives the relationship between attributes where negative and positive values show inverse and direct proportionality respectively. The corresponding rank of attributes is shown in table 9. Therefore it can be inferred that the strength of Modified Genetic Search algorithm is analyzed mathematically. The output result of PCA is shown in table 7,8,9.

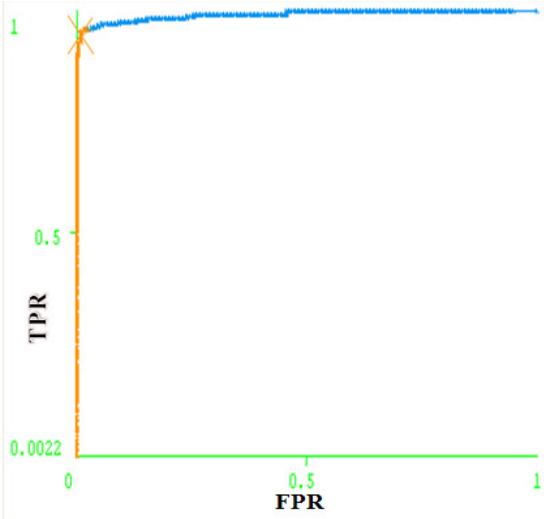

Fig 1 . ROC Area= 0.99 for case 1

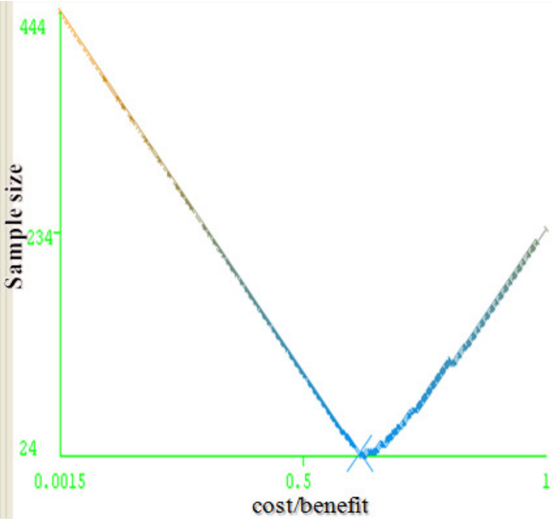

Fig 2 . Cost/benefit analysis for gain=0.7 in case 1 with 9 attributes.

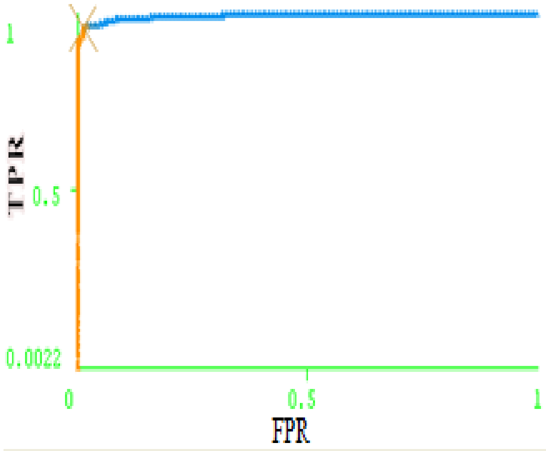

Fig 3. ROC area=0.993 for case 2

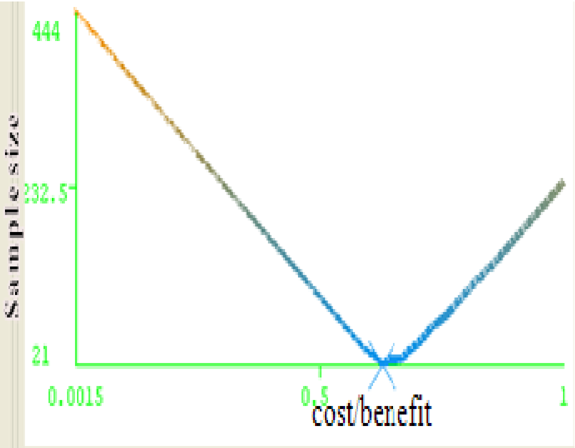

Fig 4. Cost/benefit analysis for gain=0.7 with 7 attributes in case 2

Table2: Confusion Matrix for Case 1

Predicted			
Benign	Malignant		
425	20	Benign	Actual
5	233	Malignant	

Table3: Accuracy for Case1

Accuracy by Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.955	0.025	0.986	0.955	0.97	0.99	Benign
	0.975	0.045	0.921	0.975	0.947	0.985	Malignant
Weighted Average	0.962	0.032	0.963	0.962	0.962	0.989	

Table4 Confusion matrix for Case 2

Predicted			
Benign	Malignant		
427	18	Benign	Actual
3	235	Malignant	

Table 5 Accuracy for Case 2

Accuracy by Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.96	0.021	0.988	0.96	0.974	0.993	Benign
	0.979	0.04	0.928	0.979	0.953	0.991	Malignant
Weighted Average	0.966	0.028	0.967	0.966	0.967	0.992	

Table 6:Initial population characteristics for the 20 chromosomes

Merit	Scaled	Subset
0.05417	0.05911	4 6 7 9
0.05183	0.06086	1 2 3 4 7 9
0.03953	0.07006	1 2 3 4 6 9
0.05798	0.05627	6 7 8
0.13324	0	8
0.04012	0.06962	2 3 5 6 7 8
0.04656	0.0648	2 6 7
0.09195	0.03087	5 8
0.07174	0.04598	2
0.04539	0.06568	1 6 8 9
0.04246	0.06787	3 4 5 6 7 8
0.04158	0.06853	2 4 6 7 8
0.08433	0.03656	4 5

0.06149	0.05364	2 4 7
0.04041	0.0694	1 2 4 6 7 9
0.03953	0.07006	1 3 4 6 9
0.04392	0.06677	3 6 7 8 9
0.05564	0.05802	2 4 8
0.05417	0.05911	1 4 7 8
0.04392	0.06677	3 6 7 8 9

Table 7: Correlation matrix of data set.

1	0.64	0.65	0.49	0.52	0.59	0.56	0.54	0.35
0.64	1	0.91	0.71	0.75	0.69	0.76	0.72	0.46
0.65	0.91	1	0.68	0.72	0.71	0.74	0.72	0.44
0.49	0.71	0.68	1	0.6	0.67	0.67	0.6	0.42
0.52	0.75	0.72	0.6	1	0.58	0.62	0.63	0.48
0.59	0.69	0.71	0.67	0.58	1	0.68	0.58	0.34
0.56	0.76	0.74	0.67	0.62	0.68	1	0.67	0.34
0.54	0.72	0.72	0.6	0.63	0.58	0.67	1	0.43
0.35	0.46	0.44	0.42	0.48	0.34	0.34	0.43	1

Table 8: Eigen vectors of selected attributes.

V1	V2	V3	V4	V5	V6	V7	Attributes
0.3027	-0.1424	-0.8629	-0.1024	0.0634	-0.2742	0.014	Clump_Thickness
0.3812	-0.0482	0.0153	0.2036	-0.1369	-0.0973	-0.1995	Cell_Size_Uniformity
0.3777	-0.0848	-0.0378	0.1719	-0.1043	-0.0171	-0.1242	Cell_Shape_Uniformity
0.3327	-0.0439	0.4251	-0.4651	0.0138	-0.6797	0.1256	Marginal_Adhesion
0.3363	0.1659	0.1061	0.3925	-0.6708	0.0426	0.1763	Single_Epi_Cell_Size
0.3334	-0.2546	0.0091	-0.5347	-0.123	0.604	0.3837	Bare_Nuclei
0.3461	-0.2294	0.1954	-0.011	0.251	0.2525	-0.7048	Bland_Chromatin
0.336	0.0248	0.1255	0.4475	0.6495	0.051	0.4866	Normal_Nucleoli
0.2296	0.907	-0.0885	-0.2487	0.1275	0.1415	-0.1301	Mitoses

Table 9: Ranked attributes

Ranking	Attributes
0.3456	V1
0.2593	V2
0.1994	V3
0.148	V4
0.1058	V5
0.0718	V6
0.039	V7

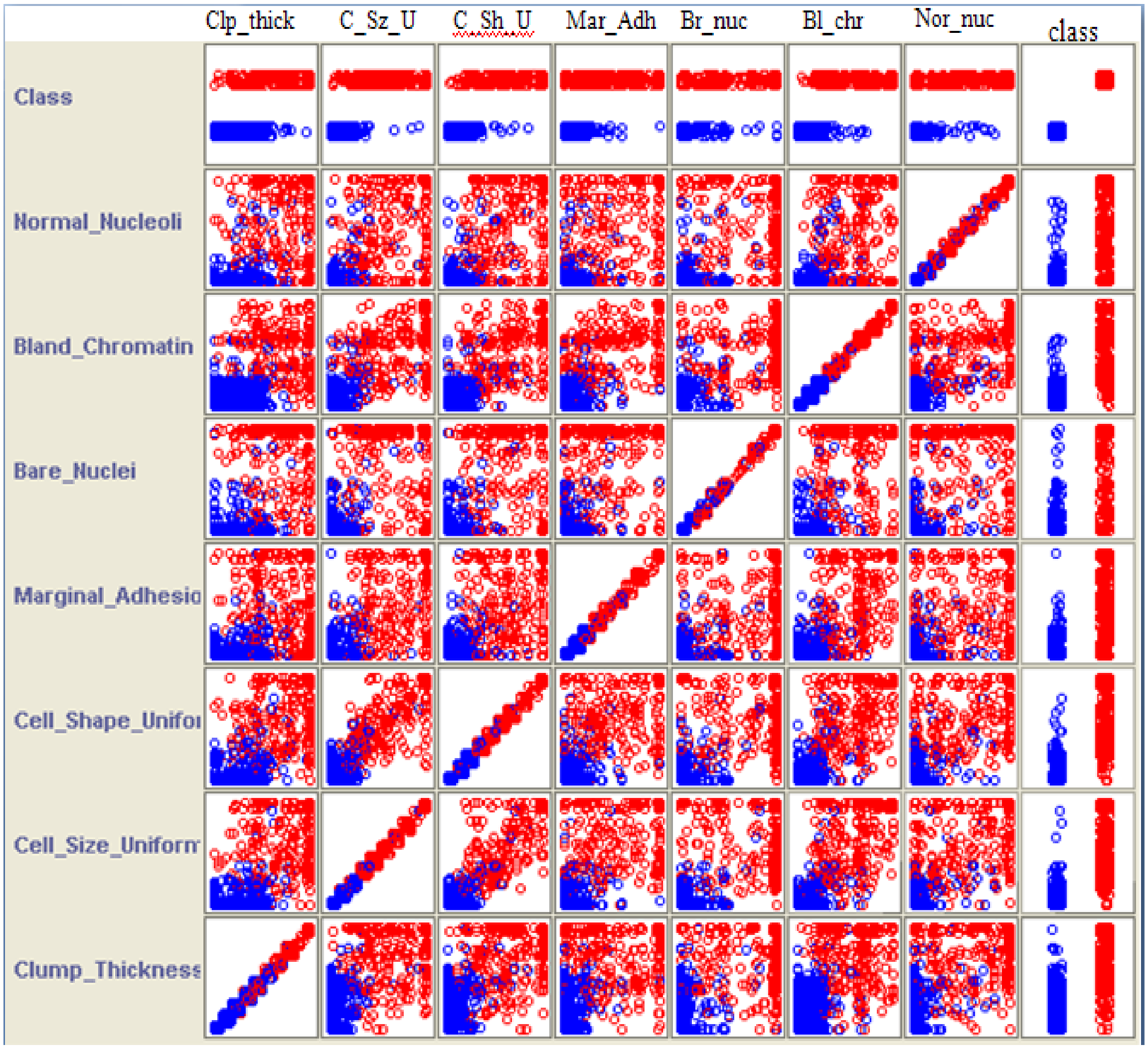

Fig5. Plot matrix distribution of selected attributes in MGSA (Class color red: Malignant, blue: Benign)

4. Conclusion

GA is a nature's inspired computational methodology to solve complex system optimal problems, and exhibit its outstanding and impressive performance especially to Nondeterministic Polynomial of some combinatorial optimizations. In this paper, the enhanced performance accuracy of 0.5856% is achieved from proposed Modified Genetic search algorithm over that of traditional model, demonstrated with a breast cancer data set. Also the results are validated mathematically using PCA. Proposed method and PCA eliminated 2 attributes from the data set. An extensive experiment has showed that improvement in initial population is effective in the optimization process. The proposed model MGSA addresses the local optima problem using merit and scaled variables in which the bad individual chromosomes are detected and deleted, reducing the search space and further iterations thereby improving efficiency. The various feature selection techniques are categorized into major category like filtering methods, wrapper subset evaluation and embedded models. The newly introduced method provided better accuracy as shown in the result section. The software reliability of Computer Aided Diagnosis system (CADx) is improved by the use of machine learning ensembles.

Although in this paper there are only nine attributes, there are still 511 possible attribute subsets that can be given as input. If the numbers of attributes increases say hundred then there are 1.27x 10^{30} possible attribute subsets to choose. In such a hard situation like this MGSA may prove helpful in determining the optimal subset. From the experimental results, it is concluded that GA is an effective algorithm that can be exploited in complex situations. It can be used in various applications such as optimization of weights of links in neural networks. Further the performance of GA can be improved by parallel implementation using MPI and other techniques and can be applied to vast field of emerging applications. Further, the methods need to be tested over larger datasets as future work.

Conflict of interest

None

References

[1] Breast cancer dataset, compiled by Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison, WI, obtained January 8, 1991.

[2] Chen Lin, “An Adaptive Genetic Algorithm Based on Population Diversity Strategy”, Third International Conference on Genetic and Evolutionary Computing. WGEC 2009. Page(s): 93 – 96.

[3] David E. Goldberg, “Genetic algorithms in search, optimization and machine learning. Addison-Wesley,1989.

[4] http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ confusion_ matrix/confusion_matrix.html

[5] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2001.

[6] Joseph F. Hair,Jr., William C.Black, Barry J. Babin, Rolph E. Anderson, Ronald L.Tatham, “ Multivariate Data Analysis”, Sixth Edition,2007, Pearson Education.

[7] Kosiński, W.; Kotowski, S.; Michalewicz, Z, “On convergence and optimality of genetic algorithms”, 2010 IEEE Congress on Evolutionary Computation (CEC), Page(s): 1 – 6.

[8] Mallik, R.K. “The Uniform Correlation Matrix and its Application to Diversity” ,2007 IEEE Transactions on Wireless Communications, Volume: 6 , Issue: 5 , Page(s): 1619 - 1625

[9] Melanie Mitchell, An introduction to Genetic Algorithms, MIT Press, Cambridge, MA,2002;first edition,1996.

[10] Nadav David Marom, Lior Rokach, Armin Shmilovici, ”Using the Confusion Matrix for Improving Ensemble Classifiers”, 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel.

[11] Nor Ashidi Mat Isa, Esugasini Subramaniam, Mohd Yusoff Mashor and Nor Hayati Othman "Fine Needle Aspiration Cytology Evaluation for Classifying Breast Cancer Using Artificial Neural Network", American Journal of Applied Sciences 4 (12): 999- 1008, ISSN 1546-9239, 2007 Science Publications

[12] Pengfei Guo, Xuezhi Wang , Yingshi Han, "The Enhanced Genetic Algorithms for the Optimization Design", 2010 IEEE 3rd International Conference on Biomedical Engineering and Informatics, pages: 2990-2994.

[13] Richard R. Picard and R. Dennis Cook, "Cross-Validation of Regression Models", Journal of the American Statistical Association, Vol. 79, No. 387 (Sep., 1984), pp. 575-583,
URL: http://www.jstor.org/stable/2288403.

[14] Ron Kohavi, George H. John (1997),"Wrappers for feature subset selection", Artificial Intelligence. 97(1-2):273-324.

[15] Sahu, S.S.; Panda, G.; Nanda, S.J., "Improved protein structural class prediction using genetic algorithm and artificial immune system" , World Congress on Nature & Biologically Inspired Computing, 2009. NaBIC 2009, Page(s): 731 – 735.

[16] Zhi-Qiang Chen, Rong-Long Wang , "An Efficient Real-coded Genetic Algorithm for Real-Parameter Optimization", 2010 IEEE Sixth International Conference on Natural Computation, pages: 2276-2280.

[17] Mandal, I., Sairam, N. Accurate Prediction of Coronary Artery Disease Using Reliable Diagnosis System (2012) Journal of Medical Systems, pp. 1-21.

[18] Mandal, I., Sairam, N. Enhanced classification performance using computational intelligence (2011) Communications in Computer and Information Science, 204 CCIS, pp. 384-391.

[19] Mandal, I. Software reliability assessment using artificial neural network (2010) ICWET 2010 - International Conference and Workshop on Emerging Trends in Technology 2010, Conference Proceedings, pp. 698-699. Cited 1 time.

[20] Mandal, I.A low-power content-addressable memory (CAM) using pipelined search scheme (2010) ICWET 2010 - International Conference and Workshop on Emerging Trends in Technology 2010, Conference Proceedings, pp. 853-858.

[21] le Cessie, S., van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.

[22] Juan J. Rodriguez, Ludmila I. Kuncheva, Carlos J. Alonso (2006). Rotation Forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(10):1619-1630.

[23] Larose, D. T. (2006) Genetic Algorithms, in Data Mining Methods and Models, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471756482.ch6

Authors

Indrajit Mandal is working as research scholar at School of Computing, SASTRA University, India. Based on his research work, he has received two Gold medal awards in Computer Science & Engineering discipline from National Design and Research forum, The Institution of Engineers (India). He has won several prizes from IITs, NITs in technical paper presentations held at national level. He has published research papers in international peer reviewed journals and international conferences. His research interest includes Machine learning, Applied statistics, Computational intelligence, Software reliability.

Sairam N is working as Professor in School of Computing, SASTRA University, India and has teaching experience of 15 years. He has published several research papers at national and international journals and conferences. His research interest includes Soft Computing, Theory of Computation, Parallel Computing and Algorithms, Data Mining.