Egyptian Computer Science Journal Vol. 37 No. 5 September 2013 ISSN-1110-2586

Improving MLP classification accuracy for breast cancer detection through evolutionary computation, partially connectivity and feature selection

[1]Smaranda Belciug, [2]Florin Gorunescu, [3]Mircea-Sebastian Serbanescu

[1]Dept. of Computer Science, Faculty of Exact Sciences, University of Craiova, Romania

[2]Dept. of Fundamental Sciences, Chair of Statistics and Computer Science, University of Medicine and Pharmacy of Craiova, Romania

[3]Dept. of Computer Science, Faculty of Mathematics and Informatics, University of Pitesti, Romania

Abstract

Breast cancer is among the deadliest cancers that kills the most people worldwide. A computer-based model to discriminate between benign and malign cases would be helpful in cancer research. We explore in this paper the feasibility of using a feature selection algorithm for removing the redundant information, and a partially connected MLP using a GA routine to find optimum weights, targeting the breast cancer detection. Experimental results obtained on two publicly-available datasets show that using a maximum relevance and minimum redundancy-based strategy for feature selection, and an evolutionary-trained partially connected MLP for patients' classification, can achieve reasonably high performance. The outcomes demonstrate that this procedure is effective, and that it is feasible to apply computational classification techniques in automatic breast cancer detection.

Keywords: *Feature Ranking and Selection, Multi-layer Perceptron, Partial Connectivity, Evolutionary Learning, Breast Cancer.*

1. Introduction

The decision process using knowledge extracted from databases represents nowadays one of the most fruitful research areas. In recent years, intelligent systems based on different machine leaning (ML) models have been intensively used in various fields, with a particular impact in the medical diagnosis process. Based on large medical databases related to different diseases, the aim of these approaches is to efficiently assist doctors in making fast and accurate diagnosis. In this context, a preliminary stage concerning special techniques for pre-processing data (e.g., feature selection) is more than desired for selecting the most relevant attributes.

Breast cancer is the second leading cause of cancer deaths in women today and one of the most common cancers among women. According to the American Cancer Society-ACS (2009) (http://www.cancer.org/), the chance of a woman having invasive breast cancer some time during her life is a little less 1 in 8, and the chance of dying from breast cancer is about 1 in 35. Fortunately, breast cancer death rates have been going down, probably due to the result of finding the cancer earlier and to better treatment.

Currently, breast cancer diagnosis is achieved using the conventional imaging (CI), or the more complex and much more expensive nuclear imaging, such as MRI (magnetic resonance imaging), PET (positron emission tomography), etc. The average accuracy of using such modern medical imaging methods for detecting breast cancer or recurrent events ranges from 80% to 90% [1], [2], [3].

From computer technology point of view, the use of evolved ML algorithms within medical decision support is now widespread and pervasive across a wide range of medical areas. In this regard, the computer-aided medical diagnosis (CAMD) became an increasingly important area for the application of the intelligent decision systems. The folded framework of a large part of these methods consists of two components. There is the feature selection module for choosing the most relevant attributes from the database, and the classifier itself consisting of a standalone algorithm or a committee of algorithms.

There are many methodologies focused on feature selection, approached from different point of views -see, for instance, [4], [5]. Regarding the use of feature selection in computer-based medical studies, one can mention association rules techniques, hill climbing algorithm, particle swarm optimization, genetic algorithms, neural networks and statistical tools [6], [7], [8], [9], [10], [11].

Various ML-based approaches have been developed as complementary techniques to support the early medical diagnosis of breast cancer. Among them, one can mention neural networks and support vector machines [12], [13], [14], [15]. In addition, there have also been some studies using hybrid neural network-genetic algorithms [16], [17], [18].

In this paper, we propose the use of a decision intelligent system, combining a previous feature selection procedure and a neural network-based classification model, designed to handle both quantitative and qualitative data, enabling the study of patterns encoded in the breast cancer data derived from two publicly-available databases. The following three questions are addressed here: (*a*) Can hybrid algorithms outperform standard neural networks regarding the breast cancer discrimination? (*b*) Which features have higher impact on the classification of breast cancer? (*c*) Can the tandem feature selection/hybrid algorithm surpass the standalone use of classification algorithms?

The remaining paper is organized as follows: Section II presents the methodology adopted in this study, followed by the description of the two datasets used for the benchmarking process depicted in section III. Section IV is devoted to the presentation of the experimental results and subsequent discussion. The paper ends with Section V, presenting the conclusions of the study.

2. Methodologies

A. Feature selection mechanism

Based on some filtering criteria, such as information gain, correlation, etc., feature selection techniques choose the most relevant attributes from the original dataset and remove unimportant or redundant features. Feature selection may be considered a key factor for

pattern recognition, since even the best classifier will perform poorly if the features are not chosen well. The feature selection methods can be broadly categorized as filters and wrappers [19]. Basically, a filter method selects features based only on the intrinsic characteristic of the data, while a wrapper method uses a search model as part of evaluation of the relevance of a feature.

Generally, a good quality feature selection technique should satisfy the following criteria [10], [20]: (*a*) "*Maximum relevance*: selected features should correlate strongest to the target variable", and (*b*) "*Minimum redundancy*: selected features should be maximally different from each other".

According to these two requirements, this study uses the following feature selection mechanism:

- A neural network is considered as classifier, and its embedded sensitivity analysis characteristic is used to rank the features in term of their predictive power. Technically, the neural network rates features through sensitivity analysis according to the deterioration in modeling performance that occurs if that feature is no longer available to the model.
- The analysis of correlation matrix to search for expected significant relations between features is then used. Since such an analysis could reveal hidden connections that may exist between the values of certain attributes, we aim to remove the redundancy among features.

The corresponding methodology is summarized in Algorithm 1.

Algorithm 1 *Feature selection*

1. Choose an appropriate neural network for a given dataset, depending on its classification performance.
2. Run the selected neural network.
3. Rank features through the sensitivity analysis provided by the neural network.
4. Calculate the correlation matrix of all features, including the decision (label) class.
5. Rank the features the most correlated with the decision class (default significance *p*-level = 0.05).
6. Find pairs of features with the highest correlation values (default significance *p*-level = 0.05).
7. If paired features are similarly ranked by the neural network and correlation matrix, remove the feature with lower rank.
8. If paired features are differently ranked, remove the feature with lower predictive power (sensitivity analysis).
9. Keep the most influential features ("core" group) achieving reasonably high classification accuracy.

Remark. It is worth noting that the issue of establishing the optimal threshold between the "core" group of the most influential attributes and the others remains an open issue, depending on the concrete decision problem to solve.

B. Partially connected hybrid MLP-GA algorithm

1. Evolutionary-trained MLP. Multi-layer perceptron (MLP) represents probably the most popular neural network architecture in use today. MLP typically consists of a set of source units that constitutes the input layer, one or more hidden layers of neurons and an output layer. A key observation in the practical use of MLP, based on a theorem due to Kolmogorov (1957), is that "a single hidden layer (i.e., a 3-layer MLP) is sufficient to uniformly approximate any continuous function in some circumstances" [21]. The training process is accomplished through continuous adjustments of the synaptic weights. The back propagation (BP) algorithm is often used to train MLP, since it still has advantages in some circumstances, and is the easiest algorithm to understand. Technically, using the gradient vector of the error surface pointing in the direction of steepest descent from the current point, one eventually finds a minimum of errors. Evaluating the derivatives of the error function, usually the *sum of squared errors* (SSE), with respect to the weights, the error function can be minimized for some optimal values of the weights [21]. Summarizing, the learning paradigm appeals to the process of adapting synaptic weights in order to minimize the network error. Accordingly, an alternative way to solve this problem is to use genetic algorithms (GAs), seen as adaptive algorithms, instead of BP. A recently developed hybrid MLP/GA algorithm has been thoroughly investigated and validated on several benchmark real-world datasets regarding the breast cancer [18]. The hybrid MLP-GA model proposed in this paper is inspired by this algorithm and consists of two components:

(*a*) The *classification* component with MLP architecture.

(*b*) The *learning* component with GA architecture.

MLP consists of a number of input units representing the number of attributes in the dataset. The network has one hidden layer with two units (one for each decision class, denoted *A* and *B*) and one output unit, representing the class attribute *A* or *B* (Fig. 1).

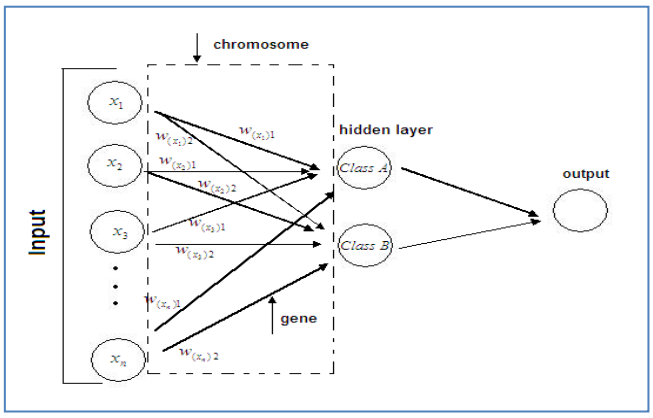

Figure (1): Architectural graph of MLP-GA

The network output is computed using the *winner-takes-all* rule. As activation function, we have used the classical sigmoid:

$$f(x) = \frac{1}{(1+e^{-x})} \tag{1}$$

A GA was designed to optimize the MLP weights, substituting the BP algorithm. The weight vector in MLP has been encoded as a vector of real numbers (network weights). Thus, a weight vector is represented through a chromosome, containing a number of genes equaling the number of neurons from the input layer multiplied by the number of neurons from the hidden layer. The weights between input units and hidden units were read off the network in a fixed order (from top to bottom) and considered as the components of the weight vector. The GA parameters have been chosen to provide optimal classification accuracy. The best performance, regardless the database, has been heuristically obtained for a population size of about 150 chromosomes and about 120 generations.

To calculate the fitness of chromosomes, the network was run on the training set and the corresponding classification accuracy (i.e., percentage of correct classified cases) has been considered. All the 150 existing chromosomes have been evaluated at each iteration of the program and the best 60 of them were kept for reproduction and mutation. Then, they have been mated with each other to replenish the population with new 60 offspring. We used the *binary tournament* selection, and, as crossover operator, we have considered the *total arithmetic recombination*, with parameter α = 0.3, experimentally chosen. The network error was directly used in the synaptic weights adjustment by taking into account the mutation operator. The worst 30 of the previously selected parents were then deliberately mutated. For each chromosome chosen for mutation, a number between 0 and 1 was randomly generated per gene. If the number is smaller than the mutation probability p_m = 0.7, experimentally chosen, then that gene is being mutated. The mutation process consisted of two steps:

(*a*) Generating a number between 0 and 1, we established whether an addition or a subtraction is being made; if the number is smaller than 0.5 then a subtraction is being made, otherwise an addition.

(*b*) Using the chromosome's error given by the formula:

$$chromosome\_error = \frac{100 - chromosome\_accuracy}{100} \tag{2}$$

each gene is being mutated according to step (*a*), subtracting or adding to the initial value the chromosome's error.

2. Partially connected hybrid MLP-GA algorithm. Starting from a standard fully-connected MLP, and based on the general manner in which a signal is being processed by the human brain, where only certain neurons participate to the course of action, a partially connected neural network can be designed by deactivating some synaptic weights. Although partially connected networks have potential advantages such as reducing training time and

less hardware requirements, there is not enough research in this area, because the building of such a simplified neural structure must take into account the specific problem to be solved. Inspired by a recently developed partially connected MLP (PCMLP), trained using the BP algorithm [22], we proposed in this paper an evolutionary-trained PCMLP, using the previously presented methodology.

The procedure to build the partially connected hybrid MLP-GA algorithm is simple and straightforward. Thus, at the beginning of the training phase we used the fully connected hybrid MLP-GA algorithm. If the weights did not suffer major modifications after a certain number t of learning samples presented to the network through the evolutionary training process (i.e., did not surpass a certain threshold heuristically estimated), then they are 'erased' from the network's architecture being inhibited (their weights being set to 0). In this study, we have experimentally considered t = 20 and threshold $\tau = 0.001$, with good computation results. Fig. 2 presents the PCMLP-GA topology (the dotted lines mean that the connection has been removed).

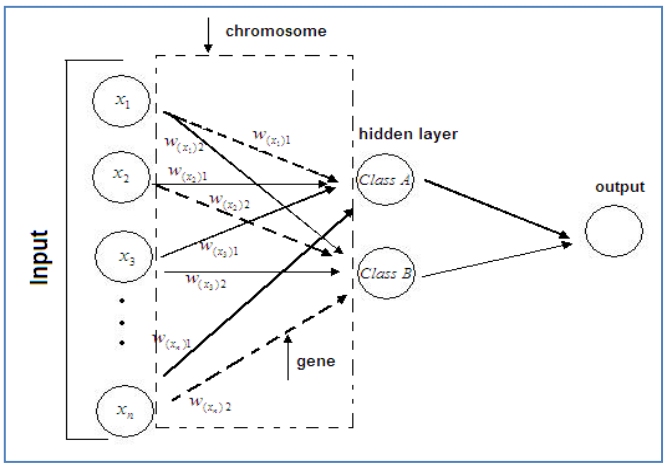

Figure (2): The PCMLP-GA scheme

Note. A Java implementation of the PCMLP-GA algorithm has been made. It is worth noting that, because JDBC (*Java Database Connectivity*) has been used, all data about patients can, at any time, be added, modified or deleted, with no change in the source of the program whatsoever.

3. Datasets Under Study

The two datasets used in this study concern breast cancer detection, and are taken from the UCI Machine Learning Repository: (http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/). The first database, the "Breast Cancer Wisconsin (Original)"-BCWO, consists of 683 cases with two decision classes: benign 444 (65%) instances and malign 239 (35%) instances. The database contains nine ordinal (categorical) attributes. The second database, the "Breast Cancer Wisconsin (Diagnostic)"-BCWD, consists of 569 cases, with two decision classes: benign 357 (62.74%) instances and malign 212

(37.25%) instances. The database contains thirty numerical attributes. For a more detailed description of these databases, readers are referred to the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

4. Results and Discussion

A. Statistical assessment

The classification performance of the tandem decision system {feature selection~PCMLP-GA} was compared to the performances of a standard MLP-BP algorithm on both the entire and the reduced datasets. Since from an engineering point of view neural networks can be thought of as "black box" machinery processing information, the statistical evaluation is essential for these experiments, bringing the benefits of improving the reliability and credibility of the obtained findings. Because both neural networks and genetic algorithms are of stochastic nature, they have to be independently run a certain number of times to obtain a reliable result regarding their robustness and effectiveness.

Based on a sample size estimation procedure (two-tailed type of null hypothesis, with default statistical power goal $P \geq 95\%$, and type I error $\alpha = 0.05$), the experiments envisaged 100 independent computer runs of each algorithm. The model validation has been achieved by using the standard 10-fold cross-validation, i.e., the entire dataset was randomly partitioned into 10 subsets of approximately equal size, and each subset in turn was used as the test set while the other subsets were used to train the classifier. The quality of classification was evaluated based on four standard statistical performance measures: *average testing accuracy* (AC), *standard deviation* (SD), *area under ROC curve* (AUC), and *confusion statistics* (CS).

The testing classification performances have been statistically compared using the *t*-test for independent samples. It is worth mentioning that the *t*-test for independent samples can be used as long as the variables are normally distributed within each group, and the variances of the two groups are not significantly different. Since the samples size, equaling 100, is fairly large, the deviation from normality observed in data does not matter much, and, on the other hand, in practice, unequal variances of two independent samples are less problematic when the samples have the same size.

B. Experimental results (original datasets)

Results of four experiments regarding the breast cancer detection are reported in this subsection. More specifically, both the novel PCMLP-GA algorithm and a classical 3-layer MLP trained with the BP algorithm (MLP-BP) have been included in the competition, being applied on the original datasets. The Statistica Neural Networks (SNN) (StatSoft. Inc., Tulsa, OK 74104, USA) package v.4. has been used for (MLP-BP). The prediction results, averaged over 100 independent computer runs, are shown in Tables (1) and (2).

Table (1): Models classification performance (BCWO)

Model	Mean testing performance measures		
	AC (%)	*SD(%)*	*AUC*
PCMLP-GA	90.71	5.65	0.978
MLP-BP	95.52	2.18	0.996

Table (2): Models classification performance (BCWD)

Model	Mean testing performance measures		
	AC (%)	*SD(%)*	*AUC*
PCMLP-GA	93.87	2.64	0.994
MLP-BP	91.54	7.57	0.992

A closer examination of the results presented in these tables reveals that:

- The classification performance of the two models differs, depending on dataset. While MLP-BP handles more efficiently categorical attributes, PCMLP-GA better performs when applied on numerical attributes. Compared with the standard MLP model, PCMLP-GA exhibits a higher accuracy on BCWD dataset and a lower accuracy on BCWO dataset.
- Compared to the performance of other neural networks, namely radial basis function (RBF) and probabilistic neural network (PNN), applied on the same datasets and reported in literature [18], PCMLP-GA had better performances. Thus, RBF provided 89.36% on BCWO and 87.42% on BCWD, while PNN provided 74.62% and 71.08% respectively.

Examples of confusion statistics, obtained in a complete cross-validation cycle of a computer run regarding the PCMLP-GA algorithm applied on both BCWO and BCWD, providing a similar performance to the mean testing performance, are displayed in Table (3) and (4). The tables show the summary statistics associated with the confusion matrix corresponding to these experiments (cases of each class that were correctly (and incorrectly) classified, and cases of that class which could not be classified at all (unknown cases)).

Table (3): Classification statistics (BCWO)

Summary	Testing classification	
	Decision class A	*Decision class B*
Correct (%)	91.89	89.95
Wrong (%)	8.11	10.05
Unknown (%)	0	0

Table (4): Classification statistics (BCWD)

Summary	Testing classification	
	Decision class A	***Decision class B***
Correct (%)	93.25	89.95
Wrong (%)	6.75	10.05
Unknown (%)	0	0

Comparing the results from the two tables above, one can see that, while the malign cases has been similarly classified, the benign cases have been better identified when using numerical attributes.

C. Feature ranking and selection

In order to assess the relevance of each individual feature derived from the two databases, we used the sensitivity analysis provided by using MLP-BP as classifier to rank the 9 features (BCWO dataset) and the 30 features (BCWD), in terms of their predictive power. The most important features in each case, ranked according to their specific influence on accuracy, are presented in Table (5).

Table (5): Feature ranking (sensitivity analysis)

Dataset	Feature ranking based on predictive power
BCWO	A6, A8, A2, A3, A5
BCWD	A16, A2, A11, A13, A20, A22, A21, A23, A24, A8, A7, A29, A27, A28, A25

Because the features represent attributes, we denoted them by A1, A2,…, regardless the dataset, without danger of confusion. The MLP-BP classification performance, namely 95.87% for BCWO, and 91.39% for BCWD, respectively, was in accordance with the mean performance obtained on each dataset.

Based on the correlation matrix, we highlighted the features the most correlated with each other, on the one hand, and with the outcome variables, namely the decision class (diagnosis), on the other hand. Since the two databases consist of different types of data -either categorical or numerical- we considered the appropriate computation approach corresponding to each type.

In the first case (BCWO dataset), consisting only of ordinal data, we used the nonparametric *Spearman R* approach, which assumes that the variables under consideration were measured on at least an ordinal scale.

Since the second dataset (BCWD dataset) consists of continuous data only, the appropriate technique is based on the standard parametric *Pearson r product-moment correlation*.

Only the significant correlations (default significance level $p < 0.05$) have been retained. Tables (6) and (7) present the correlation between the outcome and the predictors for each dataset.

Table (6): Spearman R (BCWO)

Pair of variables	Spearman R
Decision & A1	0.683
Decision & A2	0.860
Decision & A3	0.843
Decision & A4	0.737
Decision & A5	0.775
Decision & A6	0.835
Decision & A7	0.744
Decision & A8	0.748
Decision & A9	0.527

Table (7): Pearson R (BCWD)

Pair of variables	Pearson r	Pair of variables	Pearson r
Decision & A1	0.73	Decision & A21	0.78
Decision & A2	0.42	Decision & A22	0.46
Decision & A3	0.74	Decision & A23	0.78
Decision & A4	0.71	Decision & A24	0.73
Decision & A5	0.36	Decision & A25	0.42
Decision & A6	0.60	Decision & A26	0.59
Decision & A7	0.70	Decision & A27	0.66
Decision & A8	0.78	Decision & A28	0.79
Decision & A9	0.33	Decision & A29	0.42
Decision & A11	0.57	Decision & A30	0.32
Decision & A13	0.56		
Decision & A14	0.55		
Decision & A16	0.29		
Decision & A17	0.25		
Decision & A18	0.41		

The ranking process provided by both the feature selection and the correlation analysis between the decision class and the predictive attributes shows some interesting observations:

- There are some notable differences between the two ranking procedures. In the first case, while feature selection ranked attributes as {A6, A8, A2, A3, A5}, correlation analysis ranked attributes as {A2, A3, A6, A5, A8}. In the second case, feature selection provided the hierarchy {A16, A2, A11, A13, A20, A22, A21, A23, A24, A8, A7, A29, A27, A28, A25}, while correlation analysis ranked attributes as {A28, A8, A21, A23, A24, A7, A27, A11, A13, A2, A25, A29, A16}.

- Since the classification is performed by neural networks, we have decided to give a higher weight to feature selection than the correlation analysis, as presented in Algorithm 1. However, the latter can provide useful information regarding the relationship between the predictive attributes.

Since the mutually highly correlated features may not contribute to the improvement of the prediction performance, we have removed the features with the highest correlation values but with lower predictive power provided by both the sensitivity analysis and the correlation matrix, as shown in Algorithm 1. Technically, for the BCWO dataset we kept A2 only from the selected features, since the *Spearman R* correlation between A2 and A3 equaled 0.894. Accordingly, we considered for this case two reduced datasets: BCWO_1 = {A6, A8, A2, A3, A5} with 5 attributes, provided by the feature selection procedure, and BCWO_2 = {A6, A8, A2, A5} with 4 attributes, provided by both feature selection and correlation matrix analysis. In the second case (BCWD dataset), since the pairs of attributes (A16, A20), (A2, A20), (A11, A13), (A21, A8), (A21, A23), (A21, A24), (A21, A28), (A7, A27), and (A7, A28) are highly correlated (*Pearson r* correlation > 0.80), only the first item in each pair, providing higher predictive power, has been retained. Accordingly, we considered for this case two reduced datasets: BCWD_1 = {A16, A2, A11, A13, A20, A22, A21, A23, A24, A8, A7, A29, A27, A28, A25}, provided by the feature selection procedure, and BCWD_2 = {A16, A2, A11, A21, A7, A29, A25}, provided by both feature selection and correlation matrix analysis.

D. Experimental results (reduced datasets)

After applying the feature selection procedure on the original datasets, two reduced datasets have been obtained for each case. The results of the four experiments in terms of AC and AUC, averaged over 100 independent computer runs, regarding these reduced datasets are reported in Table (8).

Table (8): Models classification performance (reduced datasets)

Dataset	Model			
	MLP-BP		***PCMLP-GA***	
	AC (%)	***AUC***	***AC (%)***	***AUC***
BCWO_1	95.49	0.995	91.76	0.992
BCWO_2	96.24	0.997	92.42	0.993
BCWD_1	93.39	0.993	94.03	0.994
BCWD_2	95.80	0.997	94.96	0.994

From Table (8) we can observe that, regardless datasets and models, the classification accuracy increases when using the most influential attributes only. The statistical significance of this improvement is subsequently analyzed.

E. Statistical performance comparison

The comparison results using the *t*-test for independent samples and 100 random trials (independent computer runs) for each model, regarding both the original and the reduced datasets, and the cross-comparison between datasets, regardless their dimensions, revealed different behaviors of the two algorithms. Thus:

- There is a high statistically significant difference (p-level < 0.004) between the classification accuracies on both original datasets.
- There is a very high statistically significant difference (p-level < 0.0001) between the classification accuracies obtained on the reduced BCWO_1 and BCWO_2 datasets.
- While there is no statistically significant difference (p-level = 0.11) between accuracies on BCWD_1 dataset, the difference became high (p-level < 0.003) on BCWD_2 dataset.
- The cross-comparison involving all datasets and the MLP-BP algorithm revealed:
  - A very high statistically significant difference (p-level < 0.0001) regarding BCWO_1 vs. BCWO_2, BCWD vs. BCWD_2, and BCWD_1 vs. BCWD_2.
  - A high statistically significant difference (p-level < 0.002) regarding BCWO vs. BCWO_2 and BCWD vs. BCWD_1.
  - No statistically significant difference (p-level = 0.912) regarding BCWO vs. BCWO_1.
- The cross-comparison involving all datasets and the PCMLP-GA algorithm revealed:
  - A high statistically significant difference (p-level < 0.003) regarding BCWD vs. BCWD_2.
  - A statistically significant difference (p-level < 0.01) regarding BCWO vs. BCWO_2 and BCWD_1 vs. BCWD_2.
  - No statistically significant difference (p-level > 0.2) regarding BCWO vs. BCWO_1, BCWO_1 vs. BCWO_2, and BCWD vs. BCWD_1.

Overall, the four reduced datasets have different impacts on the two classification algorithms. Undoubtedly, accuracy is significantly improved by the prior use of the feature selection mechanism. It is worth mentioning that, although MLP-BP provided slightly better accuracy, the computation effort is much low, on the one hand, and the corresponding speed is very high, on the other hand, for PCMLP-GA.

5. Conclusion and Future Work

Neural networks have been widely used in computer-aided medical diagnosis. Among them, MLP trained with the BP algorithm represents the most extensively applied network in almost any automatic diagnosis problem, and, particularly, in breast cancer detection.

In this study, both a novel MLP version and a feature selection mechanism have been proposed to support the medical decision making in breast cancer early detection. Previous studies emphasized the improvement brought by using GA to train MLP, and the feature

selection effectiveness when used in tandem with a classifier. The current research investigates the effectiveness of using the partial connectivity to increase the MLP computational speed, the genetic algorithms to improve the training process, and a feature selection algorithm, based on both statistical and machine learning tools, to remove the redundant information.

The results obtained on two real-world databases show that relative higher prediction accuracy can be achieved by using a maximum relevance and minimum redundancy-based strategy for feature selection in tandem with a partial connected hybrid MLP-GA algorithm. Thus, the number of features dropped to half in both cases (from 9 to 5 and even 4 in the first case, and from 30 to 15 and even 7, in the second case) with a simultaneous increase of accuracy. This is not surprising, because accurate information for classification can be obtained by removing the redundant features from databases. The most significant increase in performance was achieved by MLP-BP on the reduced BCWD datasets. Although PCMLP-GA did not achieve the same classification performance, it provided a higher computation speed because of its optimal network architecture and learning paradigm.

Although this approach has proved efficient in this form, future studies should focus on improving it, regarding both the classifier and the feature selection algorithm. To conclude, the method can provide prototypes of approaches for other diseases, being ready to be applied to different situation with minor modification in structure and parameterization. Moreover, this strategy could be successfully applied to any classification problem.

Acknowledgement

The BCWO and BCWD databases were obtained from the University of Wisconsin Hospitals, Madison. Thanks go to Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian for providing the data.

References

[1] Vranjesevic, D., Filmont, J.E., Meta, J., Silverman, D.H., Phelps, M.E., Rao, J., Valk, P.E. and Czernin, J., "Whole-Body 18F-FDG PET and Conventional Imaging for Predicting Outcome in Previously Treated Breast Cancer Patients," *J Nucl Med.*, 43 (3), 325-329, 2002

[2] Gilbert, F.J., Astley, S.M., Gillan, M.G.C., Agbaje, O.F., Wallis, M.G., James, J., Boggis, C.R.M. and Duffy, S.W., "Single Reading with Computer-Aided Detection for Screening Mammography," *N. Engl. J. Med.*, 359 (16), 1675-1684, 2008

[3] Lee, J.H., Rosen, E.L., Mankoff, D.A., "The Role of Radiotracer Imaging in the Diagnosis and Management of Patients with Breast Cancer: Part 1-Overview, Detection, and Staging," *J Nucl Med.*, 50 (4), 569-581, 2009

[4] Liu, H., Motoda, H. (Eds.), "Computational methods of feature selection," Chapman and Hall/CRC, 2007

[5] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (Eds.), "Feature extraction," Foundations and applications, Springer, 2006

[6] Karabatak, M. and Ince, M.C., "An expert system for detection of breast cancer based on association rules and neural network," *Expert Syst Appl*., 36, 3465-3469, 2009

[7] Stoean, C., Stoean, R., Lupsor, M., Stefanescu, H. and Badea, R., "Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis," *Comput*. *Biol Med*., 41, 238-246, 2011

[8] Shen, Q., Meia, Z., Ye, B-X., "Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification," *Comput*. *Biol*. Med., 39, (7), 646–649, 2009

[9] Marinakis, Y., Dounias, G. and Jantzen, J., "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Comput*. *Biol. Med*., 39 (1), 69–78, 2009

[10] Zheng, H., Yang, M., Wang, H. and McClean, S., "Machine Learning and Statistical Approaches to Support the Discrimination of Neuro-Degenerative Diseases Based on Gait Analysis," in *Intelligent Patient Management*, SCI 189, 57-70, (McClean, S. et al., Eds.), Berlin Heidelberg: Springer-Verlag, 2009

[11] Gorunescu, F., Belciug, S., Gorunescu, M. and Badea, R., "Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network," *Expert Syst Appl*., 39, 12824-12833, 2012

[12] Ibeyli, E.D., "Implementing automated diagnostic systems for breast cancer detection," *Expert Syst Appl*., 33 (4), 1054-1062, 2007

[13] Kuo, S.J., Hsiao Y.H., Huang Y.L. and Chen, D.R., "Classification of benign and malignant breast tumors using neural networks and three-dimensional power Doppler ultrasound," *Ultrasound Obstet*. *Gynecol*., 32 (1), 97-102, 2008

[14] McLaren, C.E., Chen, W.P., Nie, K. and Su, M.Y., "Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques," *Acad*. *Radiol*., 16 (7), 842-851, 2009

[15] Gorunescu, F., Gorunescu, M., El-Darzi, E. and Gorunescu, S., "A statistical framework for evaluating neural networks to predict recurrent events in breast cancer," *Int*. *J Gen*. *Syst*., 39 (5), 471-488, 2010

[16] Punitha, A., Sumathi, C.P. and Santhanam, T., "A Combination of Genetic Algorithm and ART Neural Network for Breast Cancer Diagnosis," *Asian J Inform*. *Tech*., 6, 112-117, 2007

[17] El Hamdi, R., "Breast cancer diagnosis using a hybrid evolutionary neural network classifier," *Proc. 18th Mediterranean Conference on Control & Automation*, 1308 – 1315, 2010

[18] Belciug, S. and Gorunescu, F., "A hybrid neural network/genetic algorithm system applied to the breast cancer detection and reccurence," *Expert Syst*., 30 (3), 243-254, 2013.

[19] Yuan, H., Tseng, S-S., Gangshan, W. and Fuyan, Z., "A two-phase feature selection method using both filter and wrapper," *Proc. IEEE Conference on Systems, Man, and Cybernetics*, 2, 132-136, 1999

[20] Peng, H.C., Long, F. and Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, 27 (8), 1226-1238, 2005

[21] Bishop, C., "Neural networks for pattern recognition", Oxford University Press, 1995

[22] Belciug, S. and El-Darzi, E., "A partially connected neural network-based approach with application to breast cancer detection and recurrence," *Proc. 5th IEEE Conference on Intelligent Systems*, London, 191-196, 2010