**REGRESSION-BASED APPROACH FOR FEATURE SELECTION IN CLASSIFICATION
ISSUES. APPLICATION TO BREAST CANCER DETECTION AND RECURRENCE****BELCIUG Smaranda***Lecturer/Ph.D., Faculty of Mathematics and Natural Sciences /Department of Computer Science, University of Craiova, Craiova, Romania, e-mail: sbelciug@inf.ucv.ro***SERBANESCU Mircea-Sebastian***Ph.D. student, Faculty of Mathematics and Informatics/Department of Computer Science, University of Pitesti, Pitesti, e-mail: mircea_serbanescu@yahoo.com*

Abstract: Feature selection is considered a key factor in classifications/decision problems. It is currently used in designing intelligent decision systems to choose the best features which allow the best performance. This paper proposes a regression-based approach to select the most important predictors to significantly increase the classification performance. Application to breast cancer detection and recurrence using publically available datasets proved the efficiency of this technique.

Key words: classification, feature selection, regression, neural networks, breast cancer detection and recurrence

1. Introduction

Recent advances in Artificial Intelligence and Data Mining allowed researchers to develop the so-called intelligent decision models, smartly behaving like human beings. Most of them are based on a standalone machine-learning algorithm, such as artificial neural networks (NNs), genetic algorithms (GAs), support vector machines (SVMs), swarm intelligence (SI), random forests (RF), etc. [1]

The use of the intelligent decision models in medical research is a current practice, supporting the doctors' decision in a wide range of specialties. Artificial neural networks are applied to predict the severe acute pancreatitis at admission to hospital [2]. A support vector machines approach has been used in epilepsy, for the seizure prediction with spectral power of EEG [3]. Support vector machines along with linear kernel and classification trees have been used for early diagnosis of Alzheimer-type dementia [4]. A competitive/collaborative neural computing system has been considered to early detect the pancreatic cancer [5]. A hybrid neural network-genetic algorithm has been applied for the breast cancer detection and recurrence [6].

The direct use of a medical database in decision/classification purposes, without a previous analysis and pre-processing step, is often counterproductive. It is noteworthy that even the best classifier will perform poorly if the features are not chosen well. The feature selection (FS) represents the way of choosing the most relevant attributes from a dataset for improving the performance of the models used in the decision/classification process. There are many methodologies focused on FS, seen from different point of views [7-8]. The association rules technique [9], a hill climbing algorithm [10], particle swarm optimization [11], and genetic algorithms [12] are among the mostly used tools for FS in computer-aided medical diagnosis.

This study proposes a FS procedure, based on a previous regression approach, using the decision classes as the output of the linear multiple regression model, and the features as predictors. A common approach is based on the exploratory examination of the correlation matrix involving all variables, and expecting to highlight the underlying correlation between features and the decision class. However, in such an attempt, there is no "automatic" way to weed out the "false" correlations. Different from this approach, the model proposed in this paper takes advantage of the knowledge about the relationship

between several independent variables (predictors) and a dependent variable (criterion) provided by the multiple regression analysis. The degree to which the predictors (features) are related to the dependent variable (the class) is expressed in the multiple correlation coefficient R^2 .

The remaining paper is organized as follows. Section 2 describes the regression-based approach used as sensitivity analysis and providing the most relevant features for classification decision purpose, along with the two NNs proposed as classifiers. We also present the two publically available databases regarding breast cancer detection and recurrence, used to prove the consistency and effectiveness of the FS approach; the section ends with the statistical analysis benchmark. Section 3 is entirely devoted to the experimental results obtained in the concrete application to breast cancer detection and recurrence, using the regression-based FS and a standard three-layer perceptron as classifier, while Section 4 is dealing with the statistical-based assessment. The paper ends with conclusions and ideas for future research.

2. Materials and methods

FS may be considered a key factor for classification purpose, since even the best classifier will perform poorly if the features are not chosen well. It is customary to previously use FS techniques for selecting the most relevant attributes in the dataset, thus providing the elimination of unnecessary information.

2.1. Feature selection

FS implies the search for a subset of relevant attributes to be selected as being together optimally decisive for the learning task. A special tool in FS is the feature ranking, where a score is calculated for each feature according to some criterion. Based on this ranking, a certain number of features will be selected and subsequently used in the classification stage. Among statistical methods used to accomplish the FS task, we considered the regression model. It is noteworthy that the assumptions regarding the linearity (between variables) and normality (residuals) do not represent intractable issues since the multiple regression is not greatly affected by minor deviations from these assumptions. We have previously used the analysis of the correlation matrix as a first step in data analysis, searching for expected/unexpected relations between variables. The inspection of the correlation matrix is useful for establishing possible connections between the values of certain attributes for different groups of objects, thus facilitating the subsequent use of the regression model to accomplish the FS stage.

2.2. Multiple regression model

Regression models are used to predict the values of the dependent variable (outcome) from one or more other independent variables (predictors). The goal of this process is to create a model such that the predicted and observed values of the outcome variable are similar enough. Technically, the dependent variable is expressed as a (linear) combination of the explanatory variables, given by the regression equation:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_k \cdot X_{1k}, \quad (1)$$

where Y is the dependent (outcome) variable, and the variables $X_1, \dots, X_k$ are the explanatory (predictive) variables [13]. To assess the "goodness-of-fit" of the model, that is how well the model predicts the dependent variable, we consider the statistics R^2 , commonly expressed as percentage. It represents the proportion of the total sum of squares that can be explained by the regression, that is the ratio of the sum of squares due to regression and the total sum of squares provided by both regression and residuals. It is noteworthy that R^2 roughly assesses how well the model fits the data. A standard way to assess properly the "goodness-of-fit" is to analyse the residuals, by checking their 'normality', since the multiple regression assumes that the residual values (i.e., observed minus predicted values) are normally distributed.

2.3. Neural networks as classifiers

The neural network (NN) represents an information processing paradigm that is inspired by the way the human brain processes information. In this experiment, we used two well-known NNs with well established efficacy, the *multi-layer perceptron* (MLP), and the *radial basis function* network (RBF).

1. Multi-layer perceptron (MLP)

The multi-layer perceptron (MLP) consists of several layers of computing units connected together in a hierarchical feed-forward network. The main elements of a MLP are:

- input vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$, formed by p feature components x_i ;
- related output/response (multivariate) variable $y_j, j = 1, 2, \dots, q$;
- (synaptic) weights w_{ij} , connecting the output of neuron i to the input of neuron j , where neuron j lies in a layer to the right of neuron i ;
- activation non-linear function f , usually chosen sigmoidal.

A key observation in its practical use, based on the *universal approximation theorem* applied to MLP, is that a MLP with a single hidden layer (3-MLP) is sufficient to uniformly approximate any continuous function. This is the reason we have chosen a 3-MLP as classifier.

The most popular learning method for MLP is the *back-propagation algorithm* (BP). Technically, the output values are compared with the actual values and the error is computed based on a predefined error function E . According to the result obtained, one acts *backward* through the network to adjust the weights in order to minimize the error function $E(w_1, w_2, \dots, w_p)$. Under these circumstances, we want to minimize the network error function, given by:

$$SSE = \sum_{\text{training_data}} \sum_{\text{output_nodes}} (\text{actual} - \text{output})^2. \quad (2)$$

After the synaptic weights are computed in the ‘training phase’, the neural model will be used in the ‘testing phase’, thus yielding a true measure of its computing performance.

2. Radial basis function (RBF)

RBF is a symmetrical function about a center point in a multi-dimensional space. A number of hidden units with radial basis activation functions are connected in a feed-forward parallel architecture. RBF is characterized by a set of inputs and a set of outputs; between these two layers there is a hidden layer of processing units, each of them implementing a radial basis function, which has the form:

$$y_k(\mathbf{x}) = \sum_{i=1}^M w_{ki} \cdot \Phi_i(\|\mathbf{c}_i - \mathbf{x}\|, \sigma_i), \quad (3)$$

where $y_k(\mathbf{x})$ is the k -th output, $\mathbf{x}$ is an input vector, w_{ki} is the weight from the i -th kernel node to the k -th output node, $\mathbf{c}_i$ is the centroid of the i -th kernel node, σ_i is the bandwidth of the i -th kernel node, M is the number of kernel nodes, and $\Phi_i(\cdot) = \exp(-\|\mathbf{c}_i - \mathbf{x}\|^2 / (2\sigma_i^2))$ is Gaussian RBF with bandwidth of σ_i .

Note. Since RBF uses a single hidden layer, it is much faster than MLP.

2.4. Breast cancer datasets

The proposed mix model FS/MLP has been applied on two public breast cancer datasets, presented below:

1. *Breast Cancer Wisconsin (Diagnostic) -BCWD (UCI Machine Learning repository)*. BCWD consist of 569 cases, with two decision classes: benign 357 (62.74%) instances and malign 212 (37.25%) instances. Ten real-valued features are computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. For our experiment, we have considered all attributes (detailed description of the BCWD database at: <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>).
2. *Breast Cancer -BC (UCI Machine Learning repository)*. BC consists of 286 cases with two decision classes: non-recurrent-events 201 (70.27%) instances and recurrent-events 85 (29.72%) instances. The database contains nine mixed attributes, with three numerical attributes and six categorical attributes: age, tumor-size, inv-nodes, menopause, node-caps, deg-malig, breast, breast-quad, irradiat (detailed description of the BC database at:

2.5. Statistical-based evaluation of performance

Because NNs are of stochastic nature, they have been independently run 100 times to obtain a reliable result regarding both their robustness and effectiveness. Thus, from statistical point of view, we obtained an adequate statistical power (two-tailed type of null hypothesis with default statistical power goal $P \geq 95\%$, and type I error $\alpha = 0.05$).

The NN models' assessment has been achieved by using the standard 10-fold cross-validation. MLP and RBF have been independently executed 100 times (each model has been run 100 times in a complete 10-fold cross-validation cycle), and the average testing accuracy has been considered as the decision performance. Recall that the 10-fold cross-validation consists of the following steps. The classification accuracy (i.e., the percentage of correctly classified cases) in the testing stage is computed 10 times, each time leaving out one of the sub-samples and using that sub-sample as a test sample for cross-validation. Thus, each sub-sample is used 9 times as training sample and just once as testing sample. The model correct classification rates computed for each of the 10 runs of each model are then averaged to give the 10-fold estimate of the classification accuracy. This procedure is repeated 10 times to complete a cross-validation cycle, consisting of the 10 runs of the model.

The evaluation process has been assessed by statistical means, seen as an objective quality procedure. Technically, the statistical evaluation involved two comparison tests concerning the classification performances:

- Parametric t -test for independent samples;
- Difference between two proportions ($DBTP$) -two-sided test.

Recall in this context that, on the one hand, since the samples size (100) is fairly large, the deviation from normality observed in the 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. Accordingly, the use of the t -test for independent samples is thus feasible.

3. Results

3.1. Breast Cancer Wisconsin (Diagnostic)

From the correlation matrix analysis we obtained that a number of 27 independent variables (features) out of the total 30 are statistically significant correlated (p -level < 0.05) with the decision classes.

Next, applying the regression-based FS, we obtained that just 5 features really determine the decision classes (p -level < 0.05), being thus kept in model ($R^2 = 0.77$). They are radius, area, smoothness, compactness, and concavity. Taking into account that $R^2 = 0.77$, we have explained 77% of the original variability by the regression model.

The result of the two experiments, one with the entire dataset, and the other with the reduced dataset are described in Table 1.

Table 1: Testing performance/SD (%) and CPU time -BCWD dataset

Dataset	Testing accuracy/SD (%)		CPU time	
	3-MLP	RBF	3-MLP	RBF
BCWD complete	94.69/8.28	96.93/1.47	2'47"	6"
BCWD reduced	93.41/8.05	94.70/1.54	1'16"	4"

From Table 1 we can observe that, irrespective of the dataset dimension and NN type, the classification performances are fairly similar. Let us remark the notable difference in CPU time both between NN types, but, especially, when taking into account the datasets dimensions. Thus, when using the FS procedure, although the performance remains unchanged, the running time decreases dramatically.

3.2. Breast Cancer

From the correlation matrix analysis we obtained that a number of 5 independent variables (features) out of the total 9 are statistically significant correlated (p -level < 0.05) with the decision classes.

Next, applying the regression-based FS, we obtained that one feature could determine the decision classes ($R^2 = 0.16$). This feature represents the breast position: left or right. Taking into account that $R^2 = 0.16$, we have explained 16% of the original variability by the regression model. However, although a small percentage of the original variability was explained by the model, we kept only that variable in model.

The result of the two experiments, one with the entire dataset, and the other with the reduced dataset are described in Table 2.

Table 2: Testing performance/SD (%) and CPU time -BC dataset

Dataset	Testing accuracy/SD (%)		CPU time	
	3-MLP	RBF	3-MLP	RBF
BC complete	64.64/5.18	61.83/6.44	49"	4"
BC reduced	64.52/1.31	60.48/11.69	18"	2"

From Table 2 we observe that in this case too, the classification performances are fairly similar, irrespective of the dataset dimension and NN type. The same significant decreasing of the CPU time is accompanied by no change regarding the classification performance.

4. Statistical assessment

The results obtained using the two comparison tests for both the standard and reduced datasets are displayed in Table 3.

Table 3: Comparing testing performances -standard/reduced datasets

Comparison tests	t-test/ <i>p</i> -level		DBTP (<i>p</i> -level)	
	3-MLP	RBF	3-MLP	RBF
Standard BCWD vs. reduced BCWD	1.55/0.12	10.45/0.00	0.70	0.43
Standard BC vs. reduced BC	0.22/0.82	1.01/0.31	0.98	0.84

As it is shown in Table 3, there is no statistically significant difference between the classification performance of the two NNs, irrespective the datasets or dimensionality, excepting the case of the *t*-test applied to compare the performance of RBF on standard/reduced BCWD datasets. To conclude, these results show that the regression-based FS proved to be efficacious, maintaining the same performance with significantly low computation costs.

5. Conclusions

A regression-based FS model is proposed for dimension reduction, used in tandem with two NNs: 3-MLP and RBF. The paper investigates the effectiveness of this FS technique, especially designed for breast cancer detection and recurrence. The design and functionality of the model for the discrimination between malignant and benign tumors, and recurrence detection attained its planned goals. It provides a significantly better CPU time (computation effort) management, while keeping the same performance level of classification.

The method can provide a prototype of approaches for other diseases, both the FS technique and NN models being ready to be applied to different situation with minor modifications.

Future work will explore the use of the logistic regression instead of the linear regression as FS tool.

Acknowledgement.

The BC database was obtained from the Institute of Oncology, University Medical Centre, Ljubljana, Slovenia. Thanks go to M. Zwitter and M. Soklic for providing the data. The BCWD database was 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.

This work was supported by the strategic grant POSDRU/159/1.5/S/133255, Project ID 133255 (2014), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007 – 2013.

6. References

1. Turban, E., Aronson, J.E., Liang, T.-P., *Decision Support Systems and Intelligent Systems*, 7th edition, Prentice-Hall of India, (2006).
2. Andersson, B., Andersson, R., Ohlsson, M., Nilsson, J., Prediction of Severe Acute Pancreatitis at Admission to Hospital Using Artificial Neural Networks, *Pancreatology*. Vol. 11, pp. 328-335, (2011).
3. Park, Y., Luo, L., Parhi, K.K., Netoff, T., Seizure prediction with spectral power of EEG using cost-sensitive support vector machines, *Epilepsia*, Vol. 52, pp. 1761-1770, (2011).
4. Salas-Gonzalez, D., Gorriz, J.M., Ramírez, J., *et al.*, Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees, *Phys Med Biol.* Vol. 55, No. 10, pp. 2807-2817, (2010).
5. Gorunescu, F., Gorunescu, M., Saftoiu, A., Vilmann, P., Belciug, S., Competitive/collaborative neural computing system for medical diagnosis in pancreatic cancer detection, *Expert Systems*, Vol. 28, No. 1, pp. 33-44, (2011).
6. Belciug, S., Gorunescu, F., A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence, *Expert Systems*, Vol. 30, No. 3, pp. 243-254, (2013).
7. Liu, H., Motoda, H., *Computational methods of feature selection*, Editors eds., Chapman and Hall/CRC, (2007).
8. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., *Feature extraction. Foundations and applications*, Editors eds., Springer, (2006).
9. Karabatak, M., Ince, M.C., An expert system for detection of breast cancer based on association rules and neural network, *Expert Systems with Applications*, Vol. 36, pp. 3465-3469, (2009).
10. Stoean, C., Stoean, R., Lupsor, M., Stefanescu, H., Badea, R., Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis, *Computers in Biology and Medicine*, Vol. 41, pp. 238-246, (2011).
11. Shen, Q., Meia, Z., Ye, B-X., Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification, *Computers in Biology and Medicine*, Vol. 39, No. 7, pp. 646-649, (2009).
12. Marinakis, Y., Dounias, G., Jantzen, J., Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Computers in Biology and Medicine*, Vol. 39, No. 1, pp. 69-78, (2009).
13. Gorunescu, F., *Data Mining: Concepts, Models and Techniques*, Springer-Verlag, Berlin Heidelberg, (2011).