Feature selection in small databases: a medical-case study

Inês Soares[1,2,*], Joana Dias[2,3], Humberto Rocha[2], Maria do Carmo Lopes[4,5], Brígida Ferreira[5,6]

[1] Department of Computer Engineering, Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal
[2] Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra), Coimbra, Portugal
[3] Faculty of Economics, University of Coimbra, Coimbra, Portugal
[4] Portuguese Institute of Oncology of Coimbra (IPOCFG, EPE), Coimbra, Portugal
[5] Institute of Nanostructures, Nanomodelling and Nanofabrication (I3N), University of Aveiro, Aveiro, Portugal
[6] School for Allied Health Technologies (ESTSP), Porto, Portugal

Abstract— **Predictions made by using machine learning classification models are recurrent in many research fields for a variety of reasons. In some cases, feature selection can efficiently improve the accuracy of classifications, while reducing the computational requirements. However, some predictive studies are characterized by a high dimensionality or based on small datasets.**

In the present paper, we apply different feature selection approaches to a small clinical dataset containing 115 patients with head-and-neck cancer treated at the Portuguese Institute of Oncology of Coimbra (IPOCFG) with Intensity Modulated Radiation Therapy (IMRT). Xerostomia is one of the most frequent long term side-effects experienced by head-and-neck cancer patients undergoing radiation therapy, reducing drastically their quality-of-life. Being able to predict xerostomia at early stages of the treatment would make it possible to adjust the treatment plan in order to minimize or avoid this complication. Different classification models to predict xerostomia are considered along with different variable screening methodologies, and the quality of each classifier is assessed by applying cross-validation procedures. The experimental results show that different variables are selected when applying different variable selection techniques with different classification models. Therefore, variable screening methods, by themselves, are not enough for predictive analysis with small datasets. Their outcome should be complemented by the incorporation of external knowledge in order to select a reduced number of both relevant and meaningful features.

Keywords— **Feature selection, small databases, classification, radiotherapy, xerostomia.**

I. Introduction

Machine learning classification models are widely used in several areas with the same purpose, the achievement of reliable inferences. For a myriad of reasons, being able to make accurate predictions is a key factor in the decision making process, whatever the study area. In a classification problem, a training dataset consisting of n elements is available. Each element is characterized by a p-dimensional attribute vector x, belonging to a suitable space, and a class label (also known as response) $y \in \{0,1,\ldots\}$. The objective is to construct a decision or classification rule (also known as predictor, classifier or model) that will accurately predict the class labels of elements for which only the attribute vector is observed. Some classification analyses are based on relatively small datasets. In such cases, the incorporation of external knowledge to the classification model construction is suggested [1]. Furthermore, some predictive studies are characterized by a high dimensionality. Indeed, in some prediction problems, a large number of potentially prognostic variables is often available. In addition of being computationally cumbersome, some features may not be plausible predictors. Some features can be highly relevant to the model, significantly improving the predictions if included and affecting negatively if removed; other features can be totally useless, since its inclusion and/or exclusion do not affect the results; and others can completely spoil the performance of the model if they are picked [2]. Therefore, in most studies, the selection of a limited number of relevant predictors is a mandatory step. In some cases, feature selection can efficiently improve the accuracy of classification, while reducing the computational requirements. Furthermore, data reduction is in concordance with the general scientific principle of parsimony, which implies that simple models are more plausible descriptions of reality than more complex ones [1]. There are a myriad of variable screening techniques that can be applied to help in the selection process of the most relevant features characterizing a specific dataset regarding the relationship with a given response [2, 3]. Nevertheless, the selection of relevant predictors is particularly problematic and challenging in small datasets.

In this paper, we present a study of the performance of different feature selection methods, from the most classic to more complex ones, applied to small datasets. A medical case study was considered and different machine learning classification algorithms to predict xerostomia radiation-induced complication for head-and-neck cancer patients irradiated with IMRT were used. Despite improvements obtained with IMRT in head-and-neck cancer patients, sparing of the salivary glands is still challenging, potentially leading to one of the most frequent long term side-effects experienced by head-and-neck cancer patients undergoing

radiation therapy – xerostomia. Therefore, being able to predict xerostomia at early stages of the radiation treatment can make it possible to adjust the treatment plan in order to minimize or avoid such complication. In the present work, different feature selection processes and predictive models were applied to predict the binary response "risk for xerostomia at 12 months of radiation treatments". A database of head-and-neck cancer patients treated at IPOCFG was used.

II. Methodological approach

A. Search Algorithms

The main objective of this paper is to illustrate the problems of using variable screening methodologies with small datasets. We focused our attention on the most used techniques, which are briefly described below.

Sequential Forward Selection (*SFS*)*:* SFS starts the search with an empty variable subset. In each step of the algorithm a new variable is selected. In the first step, all variables are considered for selection and the fitness for each variable inclusion is computed. The variable that results in the best score is included in the variable subset and excluded for the following selection steps [3]. The process is repeated until no further improvements or a pre-specified number of variables have been included.

Sequential Backward Selection (*SBS*)*:* Contrarily to SFS, the SBS algorithm starts by considering all available variables and, in each iteration, a variable is discarded. More precisely, for each possible variable removed from the data set, the fitness of the set encompassing the remaining ones is calculated and the variable that results in the best score is really excluded from the variable subset.

Sequential Forward Floating Selection (*SFFS*)*:* In SFS, once a variable is included, it cannot be excluded later [3]. The crucial idea behind SFFS is that after the inclusion of one variable the algorithm starts a backtracking phase of variables' exclusion, which is carried on until no better variables subset is found [3]. In that case, the algorithm goes back to the inclusion phase that is again followed by the backtracking exclusion phase.

Sequential Backward Floating Selection (*SBFS*)*:* Analogously to SFFS, the main idea is starting a backtracking phase of variables inclusion after the exclusion of one variable.

Genetic Algorithm (*GA*)*:* A GA is a heuristic method to search optimal solutions that mimics the process of natural evolution, by using techniques inspired on the basic genetic operators, such as inheritance, mutation, selection and crossover [2]. The GA works with populations of individuals, and each individual represents a different selection of variables. In successive generations, the population evolves and generating individuals with higher fitness values (associated with higher evaluation scores).

B. Classical Approaches

We also applied some classical variable association approaches for comparative purposes, measuring the association between each explanatory variable and the variable response. The Pearson correlation coefficient, commonly represented by *r*, is widely used in several fields to measure the degree of linear dependence between two variables. It is obtained by the ratio between the covariance of the two variables and the product of their standard deviations. Its computation results in a value belonging to [-1,1], where -1 means a perfect negative correlation, 0 no correlation and 1 a perfect positive correlation. We also performed an analysis of variance (anova) with *F*-test to measure the associations of the dependent variable with each independent but numeric variable and a chi-squared test (χ^2) for categorical variables. In this context, anova provides a statistical test to analyze the differences among variable means and the chi-squared test with 1 degree of freedom checks the independence of two variables seen as two criteria of classification of the qualitative data.

C. Classification Models

We have considered a total of six machine learning prediction models as classifiers. In the following we will describe each methodology in detail.

Random Forests (*RF*)*:* A RF consists in a collection of tree-structured classifiers [4], where leaves represent class labels and branches represent conjunctions of features that lead to those class labels. The RF classifier works as an ensemble of decision trees predictors, where each tree is constructed based on a random selection of observations of the working/training dataset. The main essence of this procedure is to build multiple trees in randomly selected subspaces of the feature space, such that locally-optimal decisions are made at each node. The split in each node is made according to the best feature among all possible features on the subspace. The classification of a new observation corresponds to the class that is the mode of the classes outputted by individual trees.

Support Vector Machines (*SVM*)*:* SVM efficiently perform a non-linear classification implicitly mapping the observations into a high-dimensional feature space using a set of mathematical functions known as kernels. The basic idea behind SVM is the construction of a hyperplane in a higher dimensional space defining a decision boundary to separate the set of elements having different class member-

ships. The algorithm selects prototypes from the training data lying on the board between two classes in order to derive the classification rule for new data [5]. SVM implementations require the user to define some parameters, namely the kernel function and a cost parameter used to penalize the classifier for incorrect classifications of the training data. The error of misclassifications can be minimized by an adequate choice of the kernel function.

Neural Networks (*NN*)*:* A NN is an interconnected group of nodes. This structure was inspired in the central nervous system and explored for addressing an array of problems [6]. Formally, a NN is an information processing paradigm composed by a large number of highly interconnected processing elements (known as neurons), organized by layers and working in unison to solve specific problems. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output layer where the answer is finally yielded. Within each hidden layer neuron there is a sigmoidal activation function that polarizes network activity, as a function of a weighted sum of its inputs, and helps it to stabilize by modifying the weights of the connections according to the input patterns to decrease the differences between the NN outputs and the true outputs of the training data. NN analysis often requires a large number of individual runs to obtain the best solution.

Model-based Clustering (*MbC*)*:* In MbC, it is assumed that all elements of the original dataset are created by a mixture of components, each described by a density function and having an associated probability or "weight" in the mixture. The class of a new element will correspond to the group defined by the mixture component that most likely created it [7]. We can adopt any probability model for components, but typically it is assumed the Gaussian finite mixture model, where each component is modeled by a single Gaussian term with the same covariance structure among classes. This procedure is well-known as Eigenvalue Decomposition Discriminant Analysis (EDDA). Furthermore, the covariance matrix can assume several parameterizations, which leads to different models with different interpretations.

K-Nearest Neighbor (*k-NN*)*:* The *k*-NN prediction rule is one of the simplest machine learning classification algorithms, being the sample neighbors taken from a set of objects, for which the class is known, and *k* a user-defined integer meaning the number of closest samples. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. To predict the class of a new instance, one first finds the *k* training samples closest to that new sample in the variable space. Then, the new sample is classified by a majority vote of its neighbors, being the sample assigned to the most common class among its *k* closest training samples [8].

Logistic Regression (*LR*)*:* The LR classifier (also known as logit model) measures the relationship between a dependent variable (also called response) and one or more independent variables, by using probability scores as the predicted values of the dependent variable. The probabilities are modeled as a function of the explanatory variables by using a logistic function [9].

D. Performance Evaluation Measures

One way of dealing with small datasets is to use bootstrapping to create the training set to be used in the classification models. This set is higher or equal in size to the original dataset generated by random selection with replacement [10]. In this study we have applied all classifiers with and without bootstrapping, in the latter case considering sets with 500 observations.

The evaluation score used was the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). For a binary outcome, ROC and AUC are the most commonly recommended and used performance measures to judge the discriminative ability of a model between the observations with and without the characteristic [11]. Several authors claim that the Area Under the Precision-Recall Curve (AUPRC) is more suitable for unbalanced datasets [11]. Therefore, this measure was also considered as fitness function in the experiments.

The original dataset composed by 115 observations was split into 2 subsets, a validation set with *N* observations and a training and testing set with the remaining 115–*N* observations, which was used for variable screening by applying the cross-validation technique incorporating or not bootstrap samples [12]. Cross-validation involves the partitioning of the available data sample into complementary subsets, performing the analysis on one subset (training set) and testing the analysis on the other subset (testing set). The algorithm was run 100 times for each variable screening approach and classification model. For those runs resulting in an AUC/AUPRC of the validation set equal or greater than 0.5, the proportion of the times that each explanatory variable was selected by the algorithm was determined.

III. Experiments

A. Database

The clinical data of each patient, comprising a number of patient features and medical registrations (such as patient and tumor characteristics, treatment details and patient re-

sponse to radiation therapy registered during the follow-up medical consultations), were exported from the electronic health information system RESPONSE [13]. Only two severity classes were considered: "1" if the patient presented xerostomia, whatever the severity; "0" otherwise. In our dataset, 72 patients presented xerostomia, belonging to class "1", and 43 belong to class "0", being complication-free.

The aim behind this study concerned the identification of the attributes expected to be highly associated with the posterior development of xerostomia induced by radiation treatments. Given that the main purpose is to be able of predicting xerostomia for new patients, the only information that can be used for variable screening is the one known at the beginning of the treatment or, at most, during the first weeks of treatment. The attributes considered in our study concern: the patient's data, the treatments applied before or concomitantly with the radiotherapy, the treatment technique, the overall planned treatment time, the severity of xerostomia prior to radiotherapy and the planned mean dose on the tumor and organs at risk. The features set is described with detail in [14].

B. Results and Discussion

The results obtained by the most classical approaches as well as the most widely used search algorithms converged to the same conclusions.

The Pearson correlations depicted on Table 1 together with the statistical significances show that no independent variable is highly correlated with the development of xerostomia. The only correlations statistically significant at the 0.05 level are with *Gender* and *treatment technique dIMRT* variables. However, the corresponding *r* values revealed a clearly weak association between those attributes and the development of xerostomia induced by radiation.

The Anova analyses performed for continuous variables show that the values taken by each explanatory variable do not differ significantly between patients with and without xerostomia (Table 1). The hypotheses for this statistical test were H_0: $\mu_1 = \mu_2$ and H_1: at least one of the means is different. If the *F* value produced by anova was greater than the critical *F-test* value (equal to 3.925), the null hypothesis was rejected; otherwise (as happened), we are not able to reject H_0 and so we cannot say that there is a relationship between features. The hypotheses for the chi-test were H_0: variables are independent and H_1: variables are dependent. If the *Q* value produced by χ^2 test is equal or greater than the *Q-test* value with 1 degree of freedom (equal to 3.841), the null hypothesis is rejected; otherwise (as happened), we are not able of rejecting H_0 and so, we cannot say that there is a relationship between features. The chi-test for categorical variables did not evidence thus a dependence of the feature values and the development of xerostomia induced by radiation (Table 1).

Table 1 Classical measures.

Explanatory variables	Pearson Correlations		Anova	χ^2
	R	*p*-value	*F*	*Q*
Xerostomia at baseline (X1)	-0.071	0.449	-	0.444
Physical Mean Dose				
Primary Tumor	0.117	0.211	1.582	-
Salivary Glands	0.097	0.302	1.077	-
Parotids	0.141	0.134	2.277	-
Oral Cavity	0.041	0.666	0.188	-
Submandibular Glands	-0.009	0.923	0.009	-
Corrected Mean Dose for a fractionation of 2 Gy				
GTV-T1	0.081	0.387	0.076	-
Salivary Glands	0.070	0.46	0.550	-
Parotids	0.107	0.255	1.309	-
Oral Cavity	0.026	0.784	0.075	-
Submandibular Glands	-0.009	0.921	0.010	-
Number Sub. Glands	-0.072	0.445	-	0.440
Age	0.021	0.827	0.049	-
Gender	**0.221**	**0.018***	-	0.018
Treatment Technique				-
IMRT (TP1)	0.102	0.279	-	0.275
dIMRT (TP2)	**-0.191**	**0.041***	-	0.041
Type of Chemotherapy				
Chemotherapy (CT1)	-0.060	0.527	-	0.522
Cisplatina (CT2)	0.160	0.089	-	0.087
Cetuximab (CT3)	-0.100	0.289	-	0.285
Type of Radiotherapy	-0.120	0.202	-	0.199
Surgery (Yes/No)	-0.168	0.072	-	0.071
Overall Treatment Time	0.041	0.667	0.186	-

* Statistical significant at the 0.05 level.

Classical approaches suggested the use of alternative methodologies since no feature revealed a strong association with the development of xerostomia radiation side-effect. Hence, we applied different variable screening approaches in order to select those features really representatives of the data and directly allied to the development of xerostomia. Several experiments have been done, but the same conclusions were reached. The results evidenced a lack of clarity and certainty to extract a variable set highly associated with the development of xerostomia 12 months after the beginning of radiation treatments. We used the packages and commands of *R* software to create the predictor models and write the feature selection approaches. We run the different search algorithms incorporating different prediction models over different combinations of assump-

tions: with and without bootstrapping samples; considering the original small and unbalanced dataset and also balancing the original dataset before applying any technique; dealing or not with dummy variables. Dummy variables are used as devices to sort data into mutually exclusive categories. They are boolean indicators, taking the values 0 and 1 to indicate the absence and presence of some categorical effect. This way, multinomial variables can be converted into a set of binomial attributes allowing for getting additional information provided by their different and independent categories. However, none approach converged for the same group of variables. Moreover, different features sets were reached for different runs of the experiments. Since the results obtained for the different experiments are similar, we will display one figure for each experiment type.

Figures 1 and 2 show the results obtained by the application of stepwise algorithms (with and without a backtracking phase – SFS and SFFS, respectively) to the different classifiers, splitting the original small and unbalanced dataset according to *N*=15, incorporating bootstrapping samples and considering (Fig. 2) or not (Fig. 1) dummy variables. The results correspond to the proportion of selections of each feature in the 100 runs that result in AUC of the validation sets equal or greater than 0.5. We chose to use the leave-one-out cross-validation (LOOCV) procedure in the variable screening phase. LOOCV uses one sample of the dataset as test data and the remaining ones as training data, such that all samples with exception of one are used to train the model that is then used to predict the class for the remaining single sample. This procedure is repeated until each element in the dataset is used once as test data. Independently of the type of variables considered, these graphs show that the stepwise algorithms are not able of produce congruent results for the different classifiers. The application of the different stepwise selection algorithms to a specific classification model, assuming *N*=15 and applying a LOOCV procedure, is shown in Fig. 3. As we can see, the stepwise search approaches do not converge for the same set of features, even considering the same prediction model. Figures 4 and 5 illustrate the proportion of the selections of each variable when running the GA algorithm during 100 runs, considering the original dataset with *N*=25 or balancing the original dataset with *N*=21, respectively, and also assuming bootstrapping samples. In order to avoid the exhaustive and time-consuming cross-validation procedure, in the variable screening we chose to use 20% of the data as test set and the remaining 80% as training set. Once again, independently of the adopted strategy, the search algorithm was not able of circumvent the discrepancy in the selected features for the different classification approaches. Considering AUPRC as fitness function, the same landscape was observed (Fig. 6). In spite of several authors considering the AUPRC more suitable for unbalanced datasets, the variable screening algorithm is still unable to define a set of features related with the development of xerostomia after radiation.

Fig. 1 SFS algorithm applied to the original dataset.

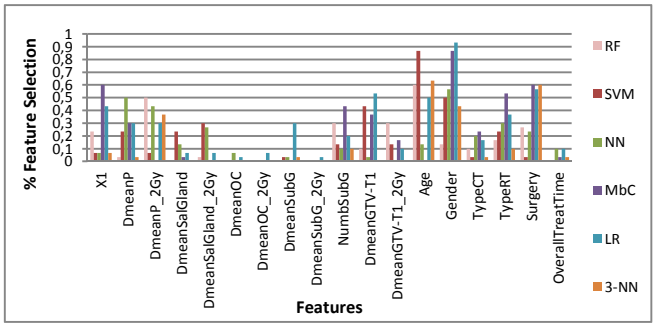


Fig. 2 SFFS algorithm by using dummy variables.

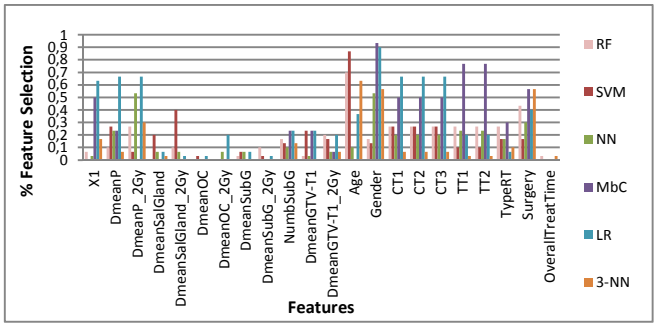
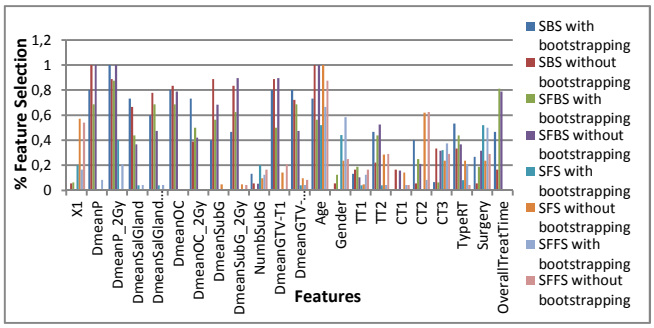


Fig. 3 Stepwise algorithms applied to the 3-NN model.



In general, the different figures (and so the different experiments) basically exhibit the same pattern. The search algorithms do not converge for the same set of features, neither considering the same prediction model with different variable screening approaches nor applying the same variable selection method to different classifiers.

IV. Conclusions

The smaller the dataset, the greater the difficulty in describing satisfactorily the patterns of the data. Consequently, different variable screening approaches will result in different variable sets whatever the machine learning prediction model used. Also, different training data sets will result in different selected attributes, and subsequently different performances, due to the small set size. The size of the dataset is undoubtedly the determinant factor in the feature selection process and consequently in predictive analyses. A small size highly affects the quality of the analyses and consequently does not produce reliable results.

Fig. 4 GA applied to the original small and unbalanced dataset.

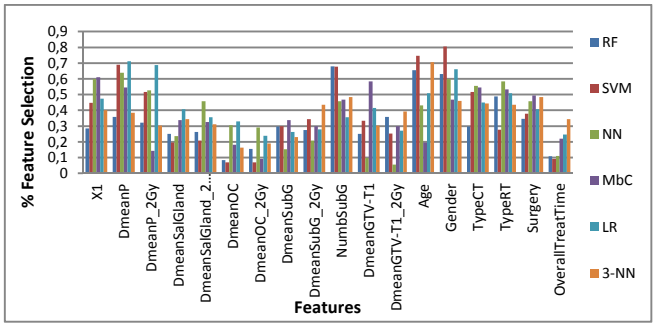


Fig. 5 GA balancing the original dataset.

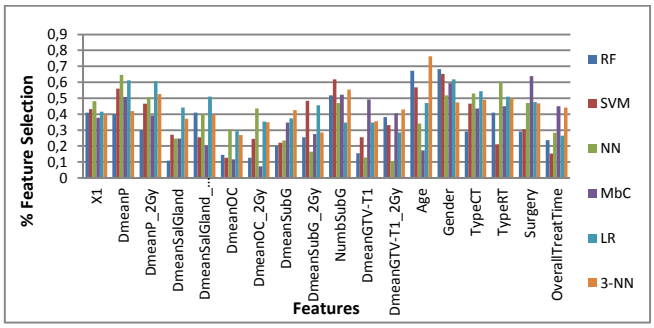
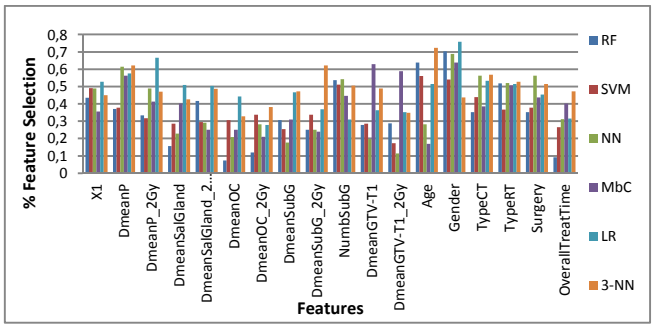


Fig. 6 GA with AUPRC as fitness function.



In conclusion, studies implemented on small datasets should incorporate as much as possible strategies alternative to the classical approaches and search algorithms for feature selection in order to guarantee reliable predictions. Moreover, their outcome should be complemented by the incorporation of external knowledge in order to select a reduced number of both relevant and meaningful features.

Acknowledgment

This work was supported by FEDER, COMPETE, iCIS (CENTRO-07-ST24-FEDER-002003), and also PTDC/EIA-CCO/121450/2010, PEst-OE/EEI/UI308/2014.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Steyerberg EW, Eijkemans MJC, Jr FEH et al. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 19:1059-79.
2. Wang L, Ni H, Yang R et al. (2013). Feature selection based on meta-heuristics for biomedicine. Optimization Methods and Software. 29:703-719.
3. Reunanen J (2003). Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research. 3:1371-1382.
4. Breiman LEO (2001). Random Forests. Mach Learn. 45:5-32.
5. Belousov AI, Verzakov SA, von Frese J (2002). A flexible classification approach with optimal generalisation performance: support vector machines. Chemometrics and Intelligent Laboratory Systems. 64:15 – 25.
6. Zhang GP (2000). Neural networks for classification: a survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews). 30:451-462.
7. Fraley C (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal. 41:578-588.
8. Schalkoff RJ. Pattern Recognition: Statistical, Structural and Neural Approaches. (Eds.) Wiley 1992 New York.
9. Bishop C. Pattern Recognition and Machine Learning. (Eds.) Springer 2006
10. Efron B, Tibshirani R. An introduction to the bootstrap. 1st CRCPress reprint. Boca Raton: Chapman & Hall/CRC 1998 237-81.
11. Davis J, Goadrich M (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning. 233 -240.
12. Molinaro AM, Simon R, Pfeiffer RM (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics. 21:3301-7.
13. Ferreira BC, Khouri L, Lopes MC et al. (2015). RESPONSE, an Electronic Health Patient Information Software for Radiation Therapy. Proceedings of the 6th Europ Conf of the Int Fed for Med and Biol Engin. 45:691-4.
14. Soares I, Dias J, Rocha H et al. (2014). Predicting Xerostomia induced by IMRT treatments: a logistic regression approach. Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference. IEEE. 72-77.