Feature selection with simple ANN ensembles

Javier Izetta and Pablo M. Granitto

CIFASIS
French Argentine International Center for Information and Systems Science
UPCAM (France) / UNR–CONICET (Argentina)
Bv 27 de Febrero 210 Bis, 2000 Rosario, República Argentina
{izetta,granitto}@cifasis-conicet.gov.ar

Abstract. Feature selection is a well-known pre-processing technique, commonly used with high-dimensional datasets. Its main goal is to discard useless or redundant variables, reducing the dimensionality of the input space, in order to increase the performance and interpretability of models. In this work we introduce the ANN-RFE, a new technique for feature selection that combines the accurate and time-efficient RFE method with the strong discrimination capabilities of ANN ensembles. In particular, we discuss two feature importance metrics that can be used with ANN-RFE: the shuffling and dE metrics. We evaluate the new method using an artificial example and five real-world wide datasets, including gene-expression data. Our results suggest that both metrics have equivalent capabilities for the selection of informative variables. ANN-RFE seems to produce overall results that are equivalent to previous efficient methods, but can be more accurate on particular datasets.

1 Introduction

Feature selection is a wide and active field of research. Two very valuable reviews are Kohavi et al. [1] and Guyon et al. [2]. Basically, feature selection is a useful pre-processing technique commonly applied to high-dimensional datasets. Its main goal is to increase the performance and interpretability of the numerical models developed on the available data, by reducing the dimensionality of the input space, discarding useless or redundant variables in an efficient way.

The introduction in the last decade of the so-called "high-throughput" technologies has created a great challenge to feature selection methods, its extension to "wide" datasets, with a high number of variables (even thousands) measured over a few samples (usually less than a hundred) [3]. Well known examples include gene expression measured with DNA microarrays [4], QSAR data [5] and mass-spectrometry applications [6, 7]. In this context, feature selection becomes highly important, because it improves the interpretability of the models, allowing the concentration of the knowledge–extraction process to a small number of variables and reducing the "black–box" effect of modern machine learning methods.

The recently introduced Recursive Feature Elimination (RFE) algorithm [8] provides good performance with moderate computational efforts when applied to wide datasets. The original and most popular version of this method uses a linear Support Vector Machine (SVM) [9] to select the features to be eliminated. This strategy is widely used in Bioinformatics [8, 10, 11] and also in Quantitative Structure Activity Relationship (QSAR) applications [5]. An alternative method was introduced by Granitto et al. [12, 13], which basically replaces SVM with Random Forest (RF) [14] into the core of the RFE method with good results.

Over the last decade ensemble methods have been on the focus of machine learning research[15, 16]. The base of these procedures is the intuitive idea that by combining the outputs of several individual predictors one might improve on the performance of a single generic one. The so-called bias/variance dilemma [17] provides formal support to the success of these strategies. According to these ideas, good ensemble members must be both accurate and diverse. Typical examples are bagging [18] and boosting [19]. Several ensemble techniques have been recently applied to artificial neural networks (ANN) [20–22]. As the diversity of ANN comes naturally from the training process randomness and from the intrinsic non-identifiability of the model, it is difficult to improve over simple strategies like using several networks trained on the same data or plain bagging [23].

In this work we combine a simple ensemble of ANNs, created with the well-known bagging method, with the efficient RFE method to produce the new ANN–RFE method for variable selection on wide datasets. We introduce two different metrics of the importance of the input variables in the ANN ensemble. One of them is based on a direct computation of the derivative of the ANN's cost function and the other in an indirect estimation using shuffled datasets, as in RF. We first demonstrate the efficiency of the method using an artificial dataset. We also evaluate the accuracy of ANN-RFE using several real-world wide datasets, comparing it with the selections made with SVMs and RF coupled with RFE.

The rest of this article is organized as follows: in Section 2, we describe the ANN–RFE feature selection scheme. In Section 3 we analyze the results of the new method, comparing it with previous results. Finally, we draw some conclusions in Section 4.

2 The ANN–RFE method

The RFE selection method [8] is a recursive process that ranks variables according to a given measure of their importance. At each iteration the importance of a set of variables is measured and the less relevant one is removed. Another possibility, which is the most commonly used, is to remove a group of features each time, in order to speed up the process. Usually, 10% of the variables are removed at each step until the number of variables reaches a lower limit, and from that point the variables are removed one at a time [24]. The recursion is needed because for some measures the relative importance of each feature can change substantially when evaluated over a different subset of features during the stepwise elimination process (in particular for highly correlated features). The (inverse) order in which features are eliminated is used to construct a final ranking. The feature selection process itself consists only in taking the first n features from this ranking.

RFE can be used with any classifier, given that a measure of variable importance can be obtained from the model. In this work we introduce two such metrics for ANN ensembles.

Several authors have suggested to use the change in the general objective function when one feature is removed as a measure of importance [1, 8]. When introducing RFE, Guyon et al. [8] explained that, for classification problems, the ideal objective function is the expected value of the error (the error rate computed on an infinite number of examples). In practice, this ideal objective is replaced by a cost function E computed on training examples only. Such a cost function is usually a bound or an approximation of the ideal objective, chosen for convenience and efficiency reasons. The change in cost function E caused by removing feature x_i is related to $\partial E/\partial x_i$. For example, the OBD algorithm [25] uses a second order approximation of $\partial E/\partial x_i$ to prune weights in ANN. For an individual ANN with SOFTMAX outputs [26], $\partial E/\partial x_i$ can be computed directly with a trivial modification of the back-propagation algorithm. Our first metric is obtained taking the average on the ensemble of the direct computation of $\partial E/\partial x_i$. We call this metric the derivative (or dE) metric.

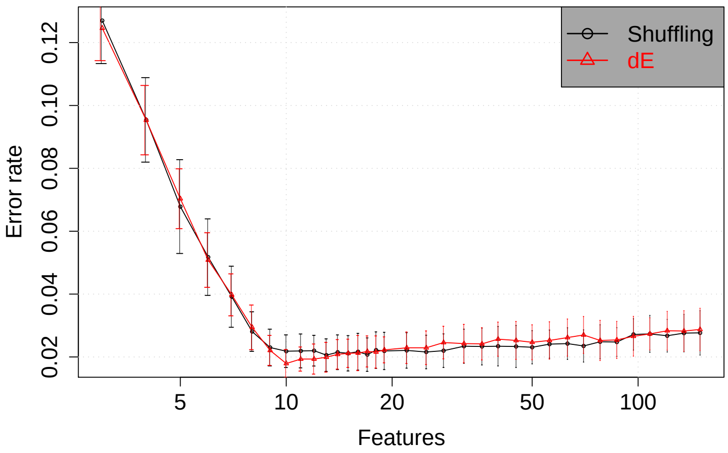
Our second method is called the *shuffling* metric. It follows the first strategy that Breiman [14] introduced to measure variable importance in RF. For any given tree in a RF there is a subset of the learning set not used by it during training, because each tree is grown only on a bootstrap sample. These subsets, called out-of-bag (OOB), can be used to give unbiased measures of prediction error. The RF shuffling metric estimates the relevance of features entering the model in the following way: one at a time, each feature is shuffled and an OOB estimation of the prediction error is made on this 'shuffled' dataset. Intuitively, irrelevant features will not change the prediction error when altered in this way, opposite to the very relevant ones. The relative loss in performance between the 'original' and 'shuffled' datasets is an indirect measure of $\partial E/\partial x_i$, and therefore correlated with the relevance of the shuffled feature. The same method can be applied straightforwardly to a bagging ensemble of ANN. For each variable and network in the ensemble we compute the difference between the original and shuffled OOB error. We then compute, for each variable, the mean error difference in the ensemble and the corresponding standard error, with which we can compute a z-score for each variable. Following Breiman again, we use the z-scores to rank the variables.

3 Empirical evaluation

3.1 An artificial example

In a first experiment we applied ANN-RFE to the "two-Gaussians" artificial dataset. In this case, even if the situation is not realistic, we know in advance which variables are relevant to the problem. We created a simple classification

Fig. 1. Mean error rates as a function of the number of variables selected by the ANN-RFE method for the artificial two-Gaussians dataset. Error bars show one standard deviation.



problem involving two classes, each one sampled from a Gaussian distribution. The problem involves 150 variables, from which 10 are relevant and the remaining 140 are uniform noise. In this and all other experiments in this work we used ANNs with a single hidden layer and SOFTMAX outputs [26]. The ANN ensembles were formed with 10 networks in this demonstrative example. The common practice in feature selection is to evaluate the performance of different methods using error curves that shows the resulting average classification error as a function of the number of variables selected. Analyzing results for some fixed number of variables is arbitrary and gives less information, and looking for the minimum error without a bias requires an additional validation set. In Figure 1 we show the classification error levels corresponding to 20 runs of the method. We show the evaluation of both versions of ANN-RFE for a complete selection process, starting with all variables and ending with subsets of 2 variables. The figure indicates that both metrics are very efficient in selecting the 10 relevant variables (dE seems to be slightly more efficient in this case).

3.2 Experimental setup

A feature selection method that uses (in any way) information about the targets may lead to overfitting, in particular with wide datasets. Thus, in order to obtain unbiased estimates of the prediction error, feature ranking and selection should be included in the modeling, and not treated as a pre-processing step; moreover, we need to appropriately decouple selection from error estimation [27].

To evaluate the methods in the real-world datasets we use a computational setup consisting of two nested processes. The outer loop performs $n = 100$ times

Table 1. Details on the five wide real-world datasets used in this work. Columns show the number of variables (V), samples (S) and classes (C), and the parameters used for the ANN: hidden units (h), learning rate (lr), momentum (m) and number of epochs (E).

Dataset	V	S	C	h	lr	m	E
Brain tumor I [11]	5921	90	5	5	0.001	0.5	500
Fragola [28]	232	233	9	8	0.0001	0.9	5000
Lampone [28]	232	92	5	10	0.0005	0.9	5000
Grana [28]	235	60	4	7	0.001	0.5	2000
Nostrani [28]	240	60	6	10	0.001	0.9	5000

a random split of the dataset in a training set (used to develop the models – including the feature selection step), and in a test set, used to estimate the accuracy of the models. The inner process supports the selection of nested subsets of features and the selection of appropriate parameters and development of classifiers over these subsets (using only the learning subset provided by the outer loop). The results of the $n = 100$ replicated experiments are then aggregated to obtain the accuracy estimation and stability evaluations.

As we stated before, we always use ANNs with a single hidden layer and SOFTMAX outputs [26]. All bagging ensembles have 100 ANN members.

3.3 Datasets

We use five different wide real-world datasets in this work. The first one corresponds to gene expression of brain tumor cells, evaluated with DNA microchips. The other four are concentration of volatiles from agro-industrial products, evaluated with PTR-MS mass-spectrometry. In Table 1 we show some details on the datasets and the original reference for each one. We also show the ANN settings we used in each case.

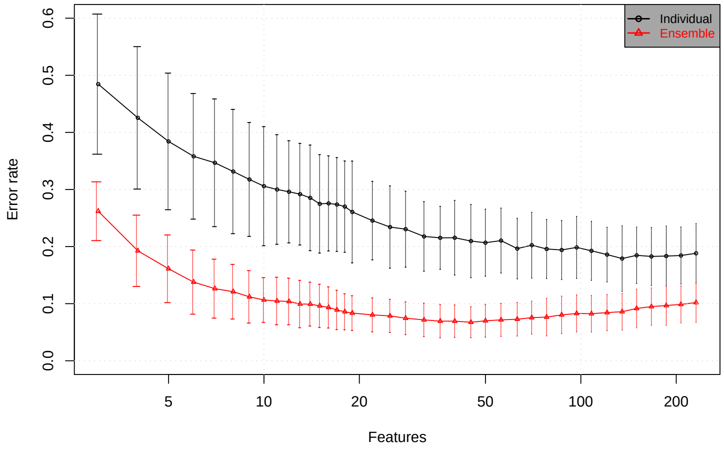
3.4 Individual ANN vs Ensembles

A relevant question at this point is if it is needed to use full ensembles to measure features importance. In Figure 2 we show a comparison between ANN-RFE applied to the Fragola dataset, using an individual ANN and an ensemble of 100 networks. Ensembles produce better classifiers, as is well known from previous works [18, 23] and also better and more stable selections, as follows from the figure. We repeated the experiment with other datasets with the same qualitative results (figures not shown for lack of space).

3.5 Evaluation of the two metrics

In Figure 3 we show a comparison of both metrics for ANN-RFE on the four PTR-MS datasets. In all four panels both metrics show a very similar performance. The error bars are also similar between methods but different among

Fig. 2. A comparison between selection made with an individual ANN and an ensemble. The figure shows error rates as a function of the number of variables selected by both methods for the Fragola dataset.



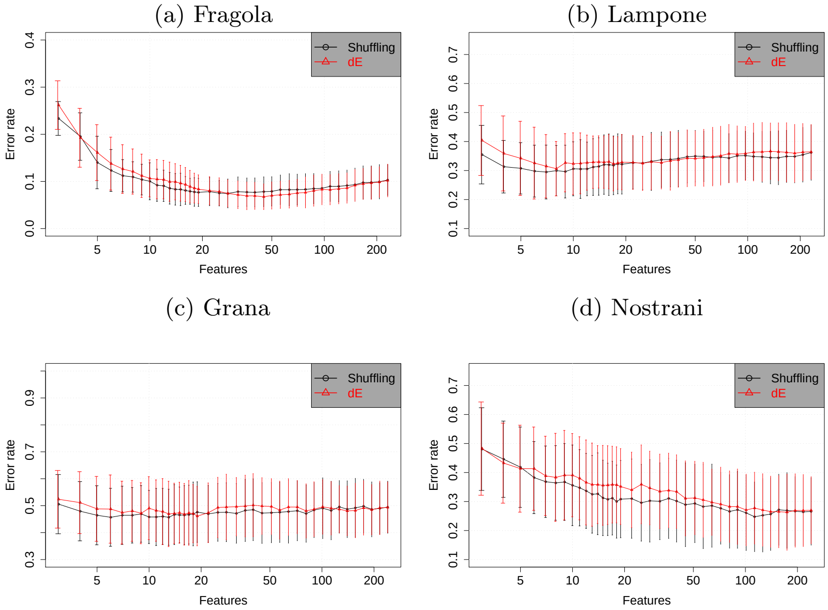
panels, because they are mostly related to the use of very small test sets in some cases. Overall, the shuffling metric shows a slightly better performance, in particular with small features subsets (the most relevant section of these experiments). In the left panel of Figure 5 we show the same comparison for the gene-expression dataset. dE shows a better performance in this case, except for subsets with very few features, when error levels have raised considerably.

3.6 Comparison with other relevance measures

In this last experiment we compared the new ANN-based metrics with three previously used measures for RFE: i) importance extracted from linear SVM [8], ii) from RF measured by shuffling the dataset [14, 13] and iii) from RF averages of the GINI index [14]. We added the RF-GINI measure because it has important similarities with our dE metric, as it is evaluated directly on the cost function using the training set only, not the OOB sets. In all cases we used the corresponding classifiers after the selection (RF for RF-based methods, SVM for SVM-based selections).

In Figure 4 we show the results of this comparison for the four PTR-MS datasets. The results are mixed, as it can be expected when comparing efficient methods. For the Fragola and Grana datasets both ANN-RFE methods show a very good performance. In both cases they seem to be slightly better than the other methods. For the Lampone dataset RF based methods are clearly superior, but the difference is based on a better discrimination more than in a better selection. In fact, ANN-RFE methods are able to reduce their classification errors when using smaller subsets, which indicates an efficient selection. In the

Fig. 3. Comparison of the two metrics introduced in this work for the ANN-RFE method. The figure shows error rates as a function of the number of variables selected by the ANN-RFE method for the PTR-MS datasets.



(a) Fragola

(b) Lampone

(c) Grana

(d) Nostrani

Nostrani dataset RF based methods performs better than ANN-RFE in the feature selection process.

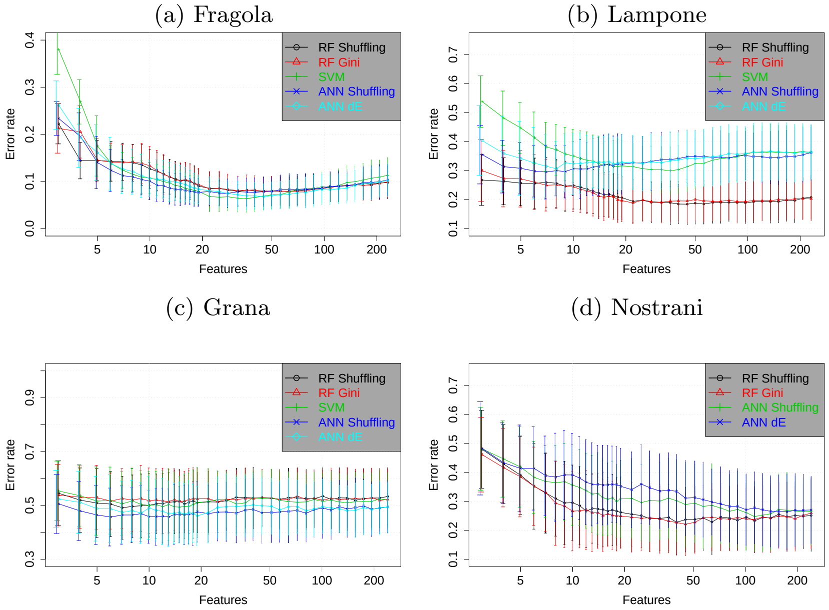
The results of the same comparison on the gene-expression dataset can be analyzed in Figure 5, right panel. This dataset shows the same situation as the Lampone dataset, but inverted. In this case ANN-RFE is always better, but the difference is more related to a better modeling than to a better feature selection.

4 Conclusions

In this paper we have introduced the ANN-RFE, a new technique for feature selection that combines the accurate and time-efficient RFE method with the strong discrimination capabilities of ANN ensembles. We also discussed two feature importance metrics that can be used with ANN-RFE: the shuffling and *dE* metrics. After showing the potential of the new method with an artificial dataset, we applied it to five real-world wide datasets.

Our results suggest that both metrics have equivalent capabilities for the selection of informative variables. Overall, ANN-RFE seems to produce results that are equivalent to previous methods. As always in machine learning, the performance of the method is highly dependent on the dataset it is being applied

Fig. 4. Comparison of different selection methods in four PTR-MS datasets. Panels show error rates as a function of the number of selected variables.


(a) Fragola

(b) Lampone

(c) Grana

(d) Nostrani

to. When faced with a problem in which ANN ensembles are the best modeling strategy (as the Brain tumors I dataset), it is expected that a feature selection strategy based directly on ANNs should give the best performance.

Several avenues are open to continue this work. Of course, a more in depth evaluation is needed, including more datasets and other aspects of the method, as for example the influence of the number of networks or the degree of overfitting of the individual members of the ensemble. The stability of the selection [28] also needs research. The dE metric has the advantage of not using at all the OOB datasets, which then could be used to produce unbiased estimates of prediction errors without keeping a test set aside, improving the efficiency of the full selection method.

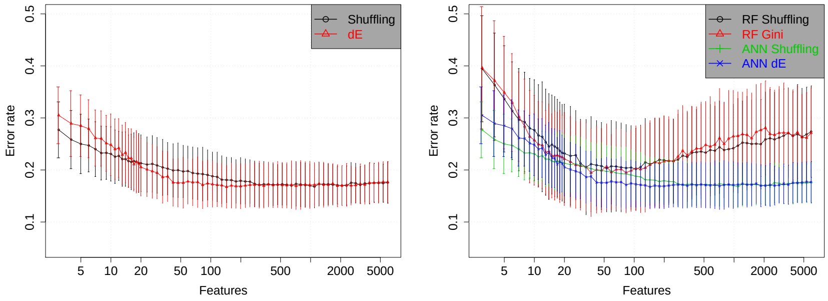
Acknowledgements

We acknowledge partial support for this project from ANPCyT (grant PICT 643).

References

1. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97, 273–324 (1996).

Fig. 5. Results on the Brain tumors I dataset (error rates as a function of the number of selected variables). Left panel: Comparison of the two metrics introduced in this work for the ANN-RFE method. Right panel: Comparison of ANN-RFE with RF-RFE.



2. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182 (2003).
3. Liu, H., Dougherty, E.R., Dy, J.G., Torkkola, K., Tuv, E., Peng, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, Z., Forman, G.: Evolving Feature Selection. IEEE Intelligent Systems, 20:6, 64–76 (2005).
4. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. Science, 286, 531-537 (1999).
5. Li, H., Ung, C. Y., Yap, C. W., Xue, Y., Li, Z. R., Cao, Z. W., Chen Y. Z.: Prediction of Genotoxicity of Chemical Compounds by Statistical Learning Methods. Chem. Res. Toxicol. 18, 1071–1080 (2005).
6. Lindinger, W., Hansel, A., Jordan, A.: On-line monitoring of volatile organic compounds at ppt level by means of Proton-Transfer-Reaction Mass Spectrometry (PTR-MS): Medical application, food control and environmental research. Int. J. Mass. Spectrom. Ion Procs. 173, 191-241 (1998).
7. Biasioli, F., Gasperi, F., Aprea, E., Mott, D., Boscaini, E., Mayr, D., Märk, T.D.: Coupling Proton Transfer Reaction-Mass Spectrometry with Linear Discriminant Analysis: a Case Study. J. Agr. Food Chem. 51, 7227-7233 (2003).
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. 46, 389–422 (2002).
9. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995).
10. Ramaswamy, S., et al.: Multiclass cancer diagnosis using tumor gene expression signatures. P. Natl. Acad. Sci. USA 98, 15149–15154 (2001).
11. Statnikov A., Aliferis, C.F., Tsamardinos, I.,Hardin, D., Levy, S.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21:5, 631–643, (2005).
12. Granitto, P.M., Biasioli, F., Gasperi, F., Furlanello, C.: Modeling Sensory Analysis datasets: the case of Italian Cheeses. Proceedings of JAIIO 2005 - The 34th International Conference of the Argentine Computer Science and Operational Research Society, Rosario, Argentina (2005).
13. Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F.: Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometr. Intell. Lab. 83, 83–90 (2006).

14. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001).
15. L. I. Kuncheva *Combining Pattern Classifiers.* Wiley-Interscience, New Jersey, 2004.
16. A. J. C. Sharkey, editor. *Combining Artificial Neural Nets.* Springer-Verlag, London, 1999.
17. S. Geman, E. Bienenstock and R. Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* 4:1-58, 1992.
18. L. Breiman. Bagging predictors. *Machine Learning* 24:123-140, 1996.
19. Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, 23-37, 1995.
20. P. M. Granitto, P. F. Verdes and H. A. Ceccatto. Neural Networks Ensembles: Evaluation of Aggregation algorithms. *Artificial Intelligence* 163:139-162, 2005.
21. B. Rosen. Ensemble learning using decorrelated neural networks. *Connection Science*. Special Issue on Combining Artificial Neural Nets: Ensemble Approaches 8(3&4):373-384, 1996.
22. H. Schwenk and Y. Bengio. Boosting neural networks. *Neural Computation* 12:1869-1887, 2000.
23. D. Opitz and R. Maclin. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, 11:169-198, 1999.
24. Furlanello, C., Serafini, M., Merler, S., Jurman, G.: Entropy-Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data. BMC Bioinformatics 4, 54 (2003).
25. Y. LeCun, J.S. Denker and S.A. Solla. Optimal brain damage. In *Advances in neural information processing systems 2, NIPS 1990*, 598-605, 1990.
26. C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, London, 1995.
27. Ambroise, C., McLachlan G.: Selection bias in gene extraction on the basis of microarray gene-expression data. P. Natl. Acad. Sci. USA 99, 6562–6566 (2002).
28. Granitto, P.M., Biasioli, F., Furlanello C., Gasperi, F.: Efficient Feature Selection for PTR-MS Fingerprinting of Agroindustrial Products. Proceedings of ICANN08, 18th International Conference on Artificial Neural Network, Prague, Czech Republic, (2008).