Inteligencia Artificial 44(2009), 27-34
doi: 10.4114/ia.v13i44.1043

INTELIGENCIA ARTIFICIAL

http://erevista.aepia.org/

Feature selection on wide multiclass problems using OVA-RFE

Pablo M. Granitto and Andrés Burgos
CIFASIS
French Argentine International Center for Information and Systems Sciences
UPCAM (France) / UNR–CONICET (Argentina)
Bv 27 de Febrero 210 Bis, 2000 Rosario, República Argentina
granitto@cifasis-conicet.gov.ar

Abstract Feature selection is a pre–processing technique commonly used with high–dimensional datasets. It is aimed at reducing the dimensionality of the input space, discarding useless or redundant variables, in order to increase the performance and interpretability of models. For multiclass classification problems, recent works suggested that decomposing the multiclass problem in a set of binary ones, and doing the feature selection on the binary problems could be a sound strategy. In this work we combined the well–known Recursive Feature Elimination (RFE) algorithm with the simple One–Vs–All (OVA) technique for multiclass problems, to produce the new OVA–RFE selection method. We evaluated OVA–RFE using wide datasets from genomic and mass–spectrometry analysis, and several classifiers. In particular, we compared the new method with the traditional RFE (applied to a direct multiclass classifier) in terms of accuracy and stability. Our results show that OVA–RFE is no better than the traditional method, which is in opposition to previous results on similar methods. The opposite results are related to a different interpretation of the real number of variables in use by both methods.

Keywords: Feature selection, multiclass, one-vs-all, wide datasets

1 Introduction

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 models developed on the dataset, by reducing the dimensionality of the input space, discarding useless or redundant variables in an efficient way. Feature selection is a wide and active field of research. Two valuable reviews are Kohavi et al. [16] and Guyon et al. [12].

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) [19]. Well known examples include gene expression measured with DNA microarrays [8], QSAR data [17] and mass-spectrometry applications [18, 3]. 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 new machine learning methods.

To work with wide datasets, the recently introduced Recursive Feature Elimination (RFE) algorithm provides good performance with moderate computational efforts [13]. The original and most popular version of this method uses a linear Support Vector Machine (SVM) [22] to select the features to be

ISSN: 1988-3064(on-line)

eliminated. This strategy is widely used in Bioinformatics [13, 20, 21] and also in Quantitative Structure Activity Relationship (QSAR) applications [17]. An alternative method was introduced by Granitto et al. [10, 11], which basically replaces SVM with Random Forest (RF) [5] into the core of the RFE method.

Most feature selection algorithms are focused on binary (two class) problems. Multiclass problems have received much less attention, because of their increased difficulty and also because some classifiers (needed for the selection) are limited to solve binary problems. Almost all available methods for feature selection on multiclass problems are *direct methods*, i.e. methods that use a selection algorithm associated to a classifier that can handle a multiclass problem directly. For example, RF is a natural multiclass algorithm with an internal (unbiased) measure of features importance[5], thus RFE plus RF [11] is a direct method. Although SVM was originally developed to deal only with binary problems, it was extended to solve multiclass problems in different ways [6, 15]. Penalized Discriminant Analysis (PDA) [14] is also a natural multiclass classifier. Combinations of RFE with these two classifiers can also be considered as direct methods.

In the last years several methods were developed to solve a multiclass problem using an appropriate combination of binary classifiers [1, 15]. The most used strategies are the "One–vs–One" and the "One–vs–ALL" (OVA). In this last case a problem with c classes is replaced with c reduced problems, each one consisting in discriminating one of the classes from all others. Associating one of these strategies with a given feature selection method we can produce what we call a *combined method* for feature selection on multiclass problems. In a very recent work, Wang et al. [23] combined OVA with some simple feature selection methods (RELIEF, mRMR) and evaluated the combined methods on some standard benchmark datasets. The authors found that selecting an independent subset of features for each OVA classifier produces better results, in terms of accuracy, than selecting the features directly on the multiclass problem. In this work we couple the OVA strategy with the efficient RFE method to produce the OVA–RFE method, a more appropriate combined method for wide datasets. We evaluated the accuracy of the new method using several real-world wide datasets and diverse classifiers (RF, SVM and PDA) to rank the features. Using the same settings we also evaluated the stability [4, 9] of the new method.

The rest of this article is organized as follows: in Section 2, we describe the OVA–RFE feature selection scheme. In Section 3 we compare the results of the new method with corresponding direct methods. Finally, we draw some conclusions in Section 4.

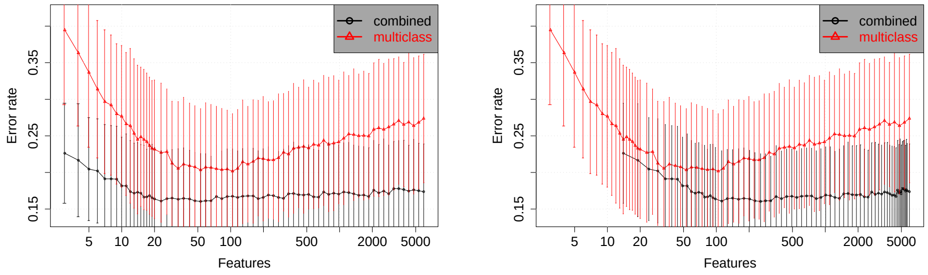
2 The OVA–RFE method

The RFE selection method [13] 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 [7]. 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 rank the features using 4 different measures: i) importance extracted from SVM [13, 11], ii) from PDA [14], iii) from RF measured by shuffling the dataset [5, 11] and iv) from RF averages of the GINI index [5].

As we stated in the introduction, the OVA method is a well-known strategy to construct multiclass classifiers starting from binary discriminant functions. In this case, a problem with c classes is decomposed into c binary problems, each one discriminating between one of the classes and all the remaining $c - 1$ classes. From this c decision functions we can estimate the posterior class probabilities and classify new examples using the Maximum-a-posteriori rule [6, 15].

The OVA–RFE method is a simple combination of RFE selection and OVA classification. In a first step, using OVA, the c–class problem is decomposed into c binary problems. Then, in the second step, we apply RFE independently to each one of the c problems, producing a potentially different rank of

Figure 1: Error rates as a function of the apparent (left panel) and real (right panel) number of variables selected by the direct (labeled multiclass, in red) and combined RF–RFE method, for the Brain Tumor I dataset.



variables for each one. In the final step, we select a given number of variables and use the c classifiers (each one fitted on its corresponding subset of features of the same length) to predict unknown samples.

3 Experiments

3.1 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 [2].

We use a computational setup consisting of two nested processes. The outer loop performs $n = 100$ times 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.

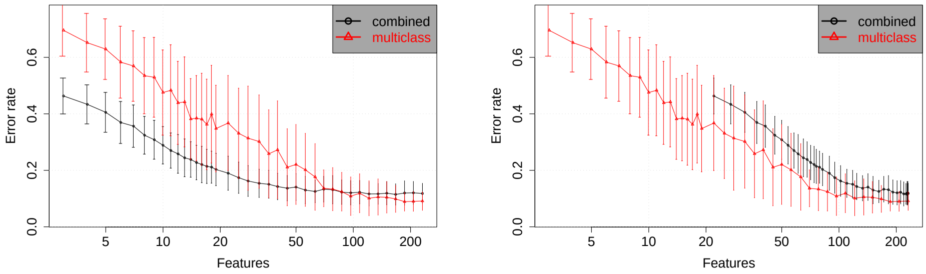
3.2 Datasets

We use 6 different wide multiclass datasets in this work. The first two correspond to gene expression 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.

Table 1: Details on the six datasets used in this work.

Dataset	Variables	Samples	Classes	Reference
Brain tumor I	5921	90	5	Statnikov et al. [21]
Brain tumor II	10367	50	4	Statnikov et al. [21]
Fragola	232	233	9	Granitto et al. [9]
Lampone	232	92	5	Granitto et al. [9]
Grana	235	60	4	Granitto et al. [9]
Nostrani	240	60	6	Granitto et al. [9]

Figure 2: Error rates as a function of the apparent (left panel) and real (right panel) number of variables selected by the direct (labeled multiclass, in red) and combined RF–PDA method, for the Fragola dataset.



3.3 Classification Errors

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 requires an additional validation set. Thus, for example in Figure 1, left panel (corresponding to RF rankings on the Brain Tumors I dataset) we show the evaluation of OVA–RFE for a complete selection process, starting with all variables and ending with subsets of 2 variables. In OVA–RFE we use the same number of variables for each one of the c binary classifiers[1]. For example, the error rates corresponding to 10 selected variables were obtained using the best 10 variables for each of the c independent binary classifiers. In the same figure we included the results of using the direct RFE method, using RF as a multiclass classifier (in all cases we show mean classification errors ± one standard deviation). The results of the new combined method seem to be clearly better than the direct method, in agreement with the results in Wang et al. [23]. The same behavior can be observed, for example, in Figure 2, left panel, corresponding to the Fragola dataset and rankings produced with PDA.

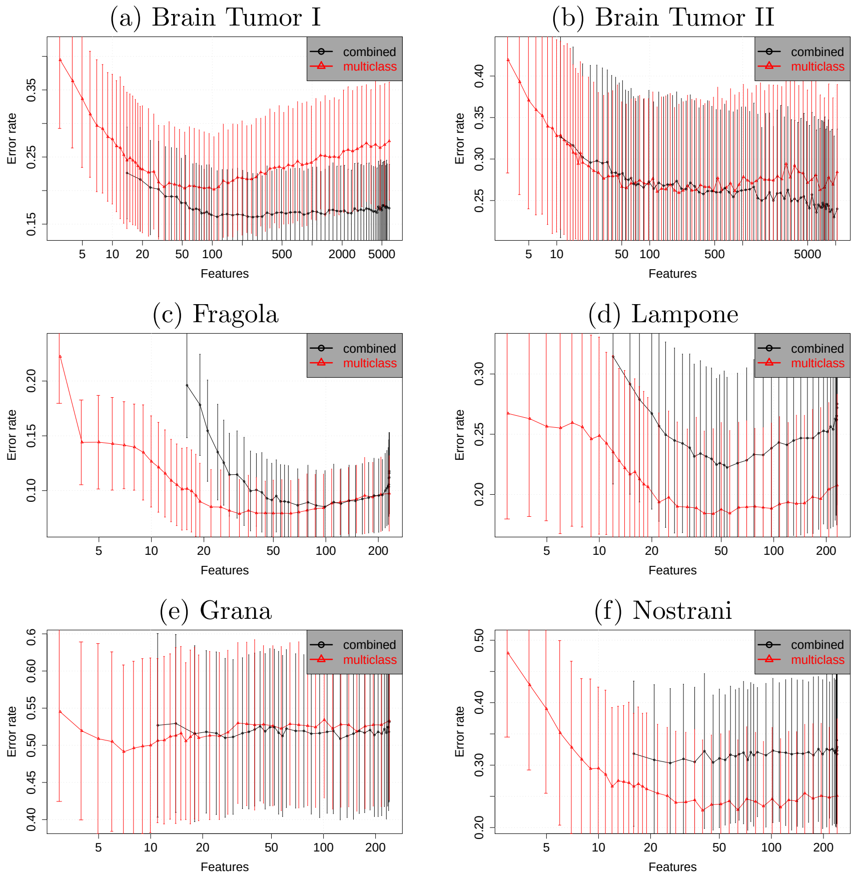
Unfortunately, that simple analysis (the same as Wang et al. [23]) is not correct. As we use an individual model for each class and we select the variables independently for each one, we can end with completely different sets of variables for each classifier. At the end, when selecting n_a variables we can get as much as $c \times n_a$ different variables used by the combined OVA classifier. We call n_a, the number of features selected at each OVA classifier, the *apparent* number of variables in the problem. The real number of variables can be obtained by counting how many different variables were selected by OVA–RFE for each n_a value. In the right panels of Figures 1 and 2 we compare the direct and combined methods but this time using the real number of variables for each algorithm. For the Brain Tumors I dataset the combined method is still better than the direct multiclass strategy, but the difference between them is considerably reduced. For the Fragola dataset the real analysis shows that the combined OVA–RFE method is no better than the direct strategy.

In Figure 3 we show a comparison of the direct multiclass method and the combined OVA–RFE for the six datasets, using the real number of variables as explained in the previous paragraph. In this case we used RF with shuffling to rank the variables. Only for the two gene expression datasets there is some gain in using the combined method. The difference between the new and traditional methods is maximized when using all the variables and becomes considerably smaller when the number of variables is reduced. This result suggests that the real difference is in the OVA strategy for classification and not in a better selection by the OVA–RFE combined method.

The *qualitative* result showed in Figure 3 are almost independent of the measure used in RFE to rank the variables. For example, in Figure 4 we show comparative results for the Fragola datasets using the four different measures we evaluated in this work, RF with shuffling, RF with Gini Index, SVM and PDA. It is clear from the figure that the four measures produce very different results when compared with each

[1] As we are analyzing full error curves, we limit ourselves to use the same number of features for each binary classifier. Of course, if one is looking only for the best performance, it is always possible to select a different number of features for each sub-problem.

Figure 3: Error rates as a function of the real number of variables selected by the direct (red) and combined (black) RF–RFE method, for the six dataset analyzed.


(a) Brain Tumor I

(b) Brain Tumor II

(c) Fragola

(d) Lampone

(e) Grana

(f) Nostrani

other, but in all cases the relation between the direct multiclass and combined methods are the same. The same property was observed for the other five datasets (figures not shown).

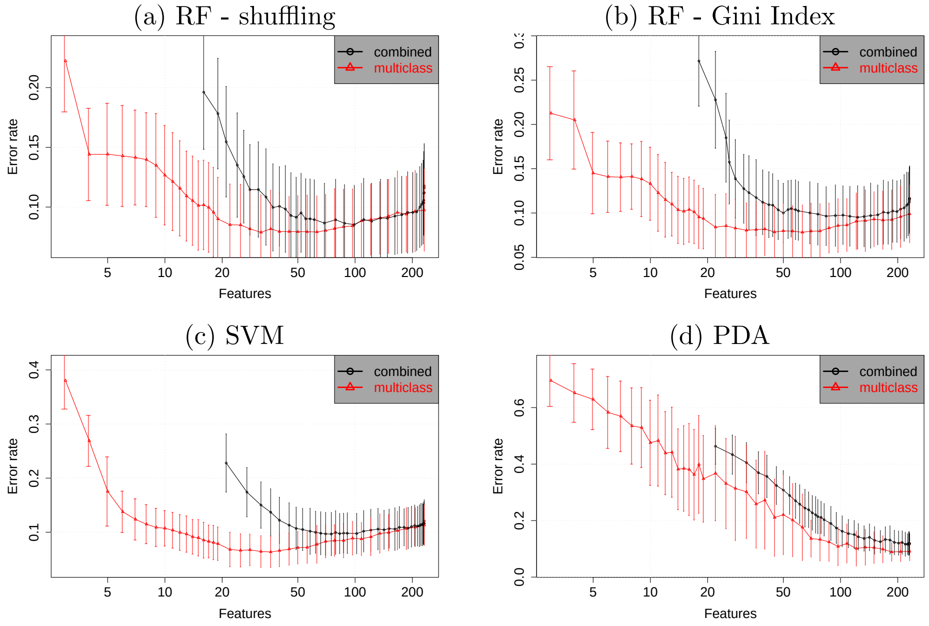
3.4 Stability

Feature selection methods are naturally unstable[4, 9] and, in consequence, each replicate of the selection process can give a different ranking. This means that the error levels showed in all previous figures are only indications of the expected behavior of the methods, but those values cannot be associated with a particular subset of variables. Of course, a higher stability of the selection method helps in the identification of the most relevant features, because rankings are more similar in that case.

To compare the stability of the direct and combined methods we assigned a relative ranking position to each feature, using a linear scale between 1 for the first position and 0 for the last one. An ideal (totally stable) selection method should return the same value for each feature in all replicates. On the opposite, a completely unstable method should return a random value in $[0 : 1]$. Thus, the dispersion of the distribution of this relative ranking (measured over the 100 replications) is correlated with the instability of the selection method.

In Figures 5 and 6 we show two examples of the analysis of the stability of the selection for both methods, direct and combined. In the figures we ordered the variables from left to right according to the mean relative ranking position, and plotted at each position a vertical line, centered at the mean value, with height equal to the corresponding standard deviation. In this kind of figure a totally stable method should produce a very thin diagonal plot from top-left to bottom-right, and the most unstable method (a

Figure 4: Error rates as a function of the real number of variables selected by the direct (red) and combined (black) method, for the Fragola dataset, using four different measures to rank the features with RFE.


(a) RF - shuffling

(b) RF - Gini Index

(c) SVM

(d) PDA

random ranking) should produce a horizontal, wide plot. In both examples (and in other cases not shown here) the direct method is clearly more stable than the combined one. This result is easy to understand, as the combined method is the result of the selection over c independent problems, and has more freedom to choice how to rank the variables.

4 Conclusions

In this paper we have introduced the OVA–RFE, a combined method for feature selection on wide multiclass datasets. Using an appropriate experimental setup, six wide datasets and four different measures of importance to rank the features we evaluated the new method and compared it with the traditional RFE version (using a direct multiclass classifier).

Wang et al. [23] showed that a combined method should work better than a direct multiclass implementation. We find the same qualitative results when evaluating OVA–RFE using Wang et al. setup. Unfortunately, those results are misleading because they are based on an incorrect way of counting the variables selected. When using the real number of variables the combined method is no better than the use of a direct multiclass classifier, considering both error levels and stability of the selections. This can be explained considering that the selections made by each OVA classifier are only “localy” optimal, focused on one class only.

Of course, the variables selected by OVA–RFE are surely better and more informative for each individual binary problem, but the objective in constructing a multiclass feature selection algorithm is to find an optimal set for the full multiclass problem. If one wants information about a given class it is easy to analyze it alone with the traditional, efficient methods.

Acknowledgements

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

Figure 5: Stability analysis for the direct (left panel) and combined (right panel) methods, using RFE plus PDA on the lampone dataset.

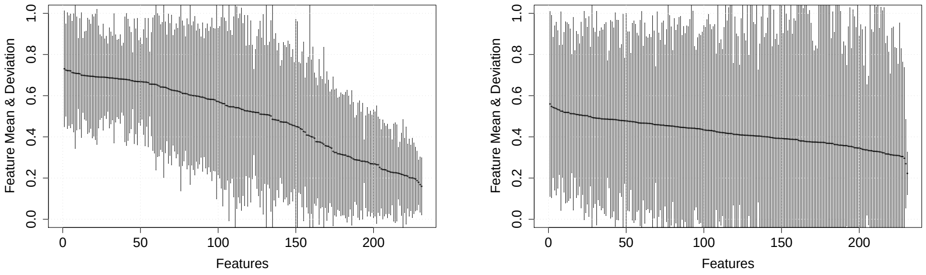
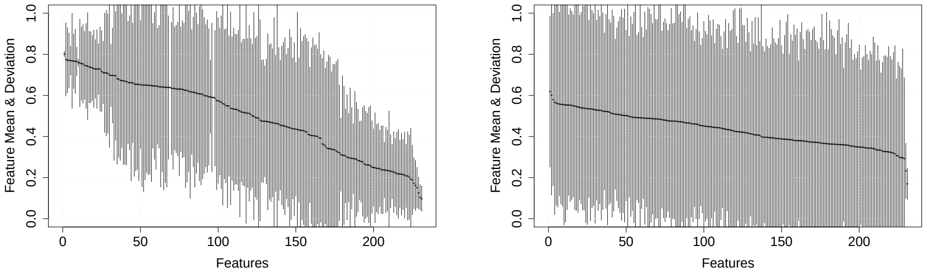


Figure 6: Stability analysis for the direct (left panel) and combined (right panel) methods, using RFE plus SVM on the fragola dataset.



References

[1] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2001.

[2] C. Ambroise and McLachlan G. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99:6562–6566, 2002. doi: 10.1073/pnas.102102699 .

[3] F. Biasioli, F. Gasperi, E. Aprea, D. Mott, E. Boscaini, D. Mayr, and T.D. Märk. Coupling proton transfer reaction-mass spectrometry with linear discriminant analysis: a case study. *Journal of Agricultural and Food Chemistry*, 51:7227–7233, 2003. doi: 10.1021/jf030248i.

[4] L Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24:2350–2383, 1996. doi: 10.1214/aos/1032181158.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324.

[6] K. Crammer and Y Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47:201–233, 2002. doi: 10.1023/A:1013637720281.

[7] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data. *BMC Bioinformatics*, (4):54, 2003. doi: 10.1186/1471-2105-4-54.

[8] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999. doi: 10.1126/science.286.5439.53.

[9] P.M. Granitto, F. Biasioli, Furlanello C., and F. Gasperi. Efficient feature selection for ptr-ms fingerprinting of agroindustrial products. In *Proceedings of ICANN08, 18th International Conference on Artificial Neural Network*, 2008. doi: 10.1007/978-3-540-87559-8.

[10] P.M. Granitto, F. Biasioli, F. Gasperi, and C. Furlanello. Modeling sensory analysis datasets: the case of italian cheeses. In *Proceedings of JAIIO 2005 - The 34th International Conference of the Argentine Computer Science and Operational Research Society*, 2005.

[11] P.M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83:83–90, 2006. doi: 10.1016/j.chemolab.2006.01.007.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. doi: 10.1.1.3.8934.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002. doi: 10.1023/A:1012487302797.

[14] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. . *Annals of Statistics*, 23:73–102, 1995. doi: 10.1214/aos/1176324456.

[15] C.-W. Hsu and Lin C.-J. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002. doi: 10.1109/72.991427.

[16] R. Kohavi and G.H John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1996. doi: 10.1016/S0004-3702(97)00043-X.

[17] H. Li, C. Y. Ung, C. W. Yap, Y. Xue, Z. R. Li, Z. W. Cao, and Chen Y. Z. Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chemical Research in Toxicology*, 18:1071–1080, 2005. doi: 10.1021/tx049652h.

[18] W. Lindinger, A. Hansel, and A. Jordan. 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. *International Journal of Mass Spectrometry and Ion Processes.*, 173:191–241, 1998.

[19] Huan Liu, Edward R. Dougherty, Jennifer G. Dy, Kari Torkkola, Eugene Tuv, Hanchuan Peng, Chris Ding, Fuhui Long, Michael Berens, Lance Parsons, Zheng Zhao, Lei Yu, and George Forman. Evolving feature selection. *IEEE Intelligent Systems*, 20(6):64–76, 2005. doi: 10.1109/MIS.2005.105.

[20] S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, C H Yeang, M Angelo, C Ladd, M Reich, E Latulippe, J P Mesirov, T Poggio, W Gerald, M Loda, E S Lander, and T R Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98:15149–54, 2001. doi: 10.1073/pnas.211566398.

[21] A Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005. doi: 10.1093/bioinformatics/bti033.

[22] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[23] L. Wang, N. Zhou, and F. Chu. A general wrapper approach to selection of class-dependent features. *IEEE Transactions on Neural Networks*, 19:1267–1278, 2008. doi: 10.1109/TNN.2008.2000395.