Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 83 (2006) 83–90
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products

Pablo M. Granitto [a,*], Cesare Furlanello [b], Franco Biasioli [a], Flavia Gasperi [a]

[a] *Istituto Agrario di S. Michele all'Adige, Via E. Mach 1, 38010 S. Michele all'Adige, Italy*
[b] *ITC/IRST Centro per la Ricerca Scientifica e Tecnologica, Via Sommarive 18, 38050 Povo, Italy*

Received 28 November 2005; received in revised form 20 January 2006; accepted 21 January 2006
Available online 9 March 2006

Abstract

In this paper we apply the recently introduced Random Forest-Recursive Feature Elimination (RF-RFE) algorithm to the identification of relevant features in the spectra produced by Proton Transfer Reaction-Mass Spectrometry (PTR-MS) analysis of agroindustrial products. The method is compared with the more traditional Support Vector Machine-Recursive Feature Elimination (SVM-RFE), extended to allow multiclass problems, and with a baseline method based on the Kruskal–Wallis statistic (KWS). In particular, we apply all selection methods to the discrimination of nine varieties of strawberries and six varieties of typical cheeses from Trentino Province, North Italy. Using replicated experiments we estimate unbiased generalization errors. Our results show that RF-RFE outperforms SVM-RFE and KWS on the task of finding small subsets of features with high discrimination levels on PTR-MS data sets. We also show how selection probabilities and features co-occurrence can be used to highlight the most relevant features for discrimination.

Keywords: PTR-MS; Feature selection; Support Vector Machines; Random forest

1. Introduction

Proton Transfer Reaction-Mass Spectrometry (PTR-MS) [1] is a relatively new spectrometric technique with a growing number of applications ranging from medical diagnosis to environmental monitoring[2]. It allows fast, noninvasive, time-continuous measurements of volatile organic compounds (VOCs). These compounds play a relevant role in food and agroindustrial applications because they are related to the real and/or perceived quality of food and to its sensory characterisation and because they are emitted during most transformation/preservation processes. Among the applications of PTR-MS based classification in food science and technology, we can cite the detection of the effect of different pasteurisation processes of fruit juices [3], the classification of strawberry cultivars [4] or the characterisation of Italian 'Grana' cheeses [5].

The output of a PTR-MS analysis consists of a spectrum with up to 500 m/z values. Each sample is associated to its PTR-MS spectrum, that is, to a vector which components are the intensities of the spectrometric peaks at different m/z ratios. Throughout this paper we will call these components 'features'. Even if the number of features (p) is relatively low compared with other spectrometric or spectroscopic approaches, in actual experiments the number of measured samples of each class (n) remains usually low, giving wide data sets ($n \ll p$) as result. Due to the absence of separation each peak in the spectrum could be related to one or more compounds. The identification of a few relevant features for the products/process under analysis is of interest for several reasons, e.g. (i) for the identification of few relevant 'quality' markers that can be measured in a simple, fast and cheap way or (ii) to concentrate to a few relevant m/z ratios the identification efforts needed to compensate for the lack of separation. While in the presented work we use PTR-MS data as an anonymous fingerprint, the possibility to identify the compounds behind the most important features is very interesting. In particular, there are indications that PTR-MS features can be related to genetic aspects of fruits [4] or to

* Corresponding author. Tel.: +39 0461 615 187; fax: +39 0461 650 957.
E-mail addresses: pablo.granitto@iasma.it (P.M. Granitto), furlan@itc.it (C. Furlanello), franco.biasioli@iasma.it (F. Biasioli), flavia.gasperi@iasma.it (F. Gasperi).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.01.007

sensory characteristics of food [6] and thus classification based on PTR-MS data could provide a tool to better investigate these fields, possibly providing a link between sensory and genetics.

Feature selection is a wide and active field of research. Valuable reviews are [7,8], and also [9,10] show application to chemometrics problems. Usually there are two different objectives for feature selection: (i) to find the subset of features with the minimum possible generalization error or (ii) to select the smallest possible subset with a given discrimination capability. In our case, the combination of the high dimensionality of PTR-MS data sets and several replications of the experiments (needed to correctly evaluate generalization error, as described later) imply prohibitive computational efforts for most selection algorithms, from exhaustive to greedy search strategies. In this context, the recently introduced Recursive Feature Elimination (RFE) algorithm provides good performance with moderate computational efforts [11]. The most popular version of this method uses a linear Support Vector Machine (SVM-RFE) to select the features to be eliminated. This strategy is widely used in Bioinformatics [11,12] and also in Quantitative Structure Activity Relationship (QSAR) applications [13,14]. Although SVM-RFE was originally developed to deal with binary problems, it can easily be extended to solve multiclass problems. We developed an alternative method [15], which basically replaces SVM with Random Forest (RF) [16] into the core of the RFE method. RF is a natural multiclass algorithm with an internal (unbiased) measure of features importance[16]. Another strategy with extended applications in bioinformatics [17] and chemometrics [18] is to rank features using some appropriate univariate statistical measure. For example, in some applications all features with no statistical significance in an ANOVA test are discarded. We selected the Kruskal–Wallis statistic (KWS) [19] in this case, because of the more general validity of its nonparametric approach [20]. In this paper we apply these three feature selection methods to PTR-MS data sets and compare their performances in the two tasks described before: to select the subsets with the minimum possible generalization error, and to select small subsets with high discrimination capabilities.

Numerous works on feature selection show biased estimation of the generalization error (the bias selection problem [21]). This is so because they use information about class labels during the selection process, which can lead to overfitting. A typical (incorrect) strategy is to perform a cross-validation loop to select the features, and then to use a new cross-validation loop, over the same samples, to estimate the test error. In this case, during the second cross-validation loop each sample set aside as test set is not completely independent of the model being evaluated, because it was previously used to select the features. More examples and discussions can be found in [21]. We use resampling methods together with replicated experiments to avoid these problems [22]. This methodology also provides partial solution to a usually neglected problem: selecting important features over replicated experiments is not straightforward, because of the instability of the selection methods [23]. Two subsets selected from slightly diverse data sets can be very different. In order to highlight which are the most important features for a given problem, we rely on the analysis of selection probability and co-occurrence of them into given subsets. We show an example illustrating which results can be obtained with these methods.

2. Experimental

2.1. Data sets

We applied the selection methods on two multiclass problems from PTR-MS analysis of agroindustrial products.

The headspace composition of the samples has been measured by direct injection in a PTRMS apparatus. Experimental details can be found in [3,4]. Each sample is associated to its PTR-MS spectrum normalised to unit total area.

The first data set considered, the Strawberry data set [4], contains 233 fruits of 9 strawberry cultivars. Classes are fairly balanced: from 21 to 30 samples per class. Each sample consists of 231 m/z values, from 20 to 250. Strawberry samples have been collected over 3 years and in different locations but here only the factor cultivar is considered for classification.

The second data set, called Nostrani [24], contains samples of 6 typical cheeses produced in cattleshed housings in Trentino, North Italy, collected during 1 year. There are 48 samples in the data set (8 from each cheese variety), each one with 240 m/z values, from 20 to 259.

2.2. Selection methods

The RFE selection method [11] is basically a recursive process that ranks features according to some measure of their importance. Fig. 1 shows the pseudo-code of the algorithm. At each iteration feature importances are measured and the less relevant one is removed. Another possibility, not used here, is to remove a group of features each time, in order to speed up the process. 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.

```
Inputs:
        Training set T
        Set of p features F = {f_1. ...., f_p}
        Ranking method M(T, F)
Outputs:
        Final ranking R
Code:
        Repeat for i in {1 : p}
                Rank set F using M(T, F)
                f* ← last ranked feature in F
                R(p − i + 1) ← f*
                F ← F − f*
```

Fig. 1. Pseudo-code for the Recursive Feature Elimination (RFE) algorithm.

Support Vector Machines are now becoming standard tools in chemometrics and QSAR modeling. They are extensively described in [25,26]. Although SVM can handle nonlinearity using kernels [27], in applications with low n/p ratios in our experience the use of linear models usually gives better results. For a problem with 2 classes $y_i \in \{-1, +1\}$, a (linear) SVM constructs a linear decision function

$$D(\boldsymbol{x}) = \text{sign}(\boldsymbol{x} \cdot \boldsymbol{w}), \tag{1}$$

where $\boldsymbol{x}$ denotes a vector with the components of a given spectra and $\boldsymbol{w}$ is a vector perpendicular to the hyperplane giving the linear decision function. According to the Structural Risk Minimization Principle [25], the SVM algorithm locates the hyperplane with the maximum margin, i.e. with the maximum distance from the hyperplane to the closest vector in each class. This leads to a constrained quadratic optimization problem that can be solved using Lagrange multipliers [25,26]. A particularity of SVMs is that the decision boundary depends only in a small subset of samples (the ones that lie closest to the hyperplane), the so-called support vectors. The components of $\boldsymbol{w}$ (w_j) are a measure of the importance of the corresponding feature to the decision function [11]. As $\boldsymbol{w}$ points in the direction of the maximum separation, a high value of a given component w_j indicates a feature j with relatively high separation between the classes.

SVMs can be extended to handle multiclass data sets using one of the various strategies for solving multiclass problems with binary classifiers [28]. Based on results from Hsu [29] we choose the One-vs.-One method. In this case, a problem with c classes is decomposed into $q = c(c-1)/2$ binary problems, each one discriminating between 2 of the c classes. To solve each problem we train a linear SVM with $C = 100$ (following [11] and empirical testing), obtaining q decision functions

$$D_i(\boldsymbol{x}) = \text{sign}(\boldsymbol{x} \cdot \boldsymbol{w}_i) \quad i = 1 \ldots q. \tag{2}$$

The weight vectors $\mathbf{w}_i$ corresponding to all binary problems are then averaged

$$\mathbf{W} = \frac{1}{q} \sum_{i=1}^{q} \boldsymbol{w}_i \tag{3}$$

and the components of $\mathbf{W}$ are used for ranking the features. Our SVM-RFE multiclass implementation thus consists in using the components of $\mathbf{W}$ as measures of feature importance for the RFE algorithm.

In a previous work [15] we introduced Random Forest-Recursive Feature Elimination (RF-RFE). Random Forest is an ensemble classification method. Basically, it grows hundreds of diverse classification trees and use them together as a composite classifier. The final classification of a given sample is decided applying the majority rule over the votes of the individual classifiers. In order to produce uncorrelated and disimilar predictions, each tree is grown using only a reduced sample (a bootstrap) of the training set. Even more, to increase the diversity between them the algorithm introduces randomness in the search of best splits [16].

Our feature ranking method is based on a measure of variable importance given by RF. For any given tree in an RF there is a subset of the learning set not used by it during training, because each tree was grown only on a bootstrap sample. These subsets, called out-of-bag (OOB), can be used to give unbiased measures of prediction error. RF 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' data set. 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' data sets is therefore related to the relevance of the shuffled feature. In RF-RFE this measure of feature importance is coupled with the RFE algorithm. As RF makes use of Out-of-Bag subsets to estimate the importances, computational efforts are not increased. Moreover, RF was developed as a multiclass algorithm, which suggests that it could provide a better measure of importance for this kind of problems than the combination of binary problems used, for example, by SVM.

We also use the nonparametric Kruskal–Wallis statistic to build univariate rankings. As KWS is a univariate measure, its value on each feature does not change when different subsets are evaluated. Clearly, recursion becomes irrelevant in this case, so only a single evaluation is needed. In this case we simply evaluate the KWS on each feature and rank them in descending value of this statistic.

2.3. Replicated experiments

Any feature selection method which uses (in any way) information about the targets can lead to overfitting, in particular with the very low n/p ratios typical of spectrometric experiments. Thus, in order to obtain unbiased estimates of the prediction error, the selection of features should be included in the modeling, and not treated as a pre-processing step, as is often made. Also, as already mentioned, most selection methods are unstable [23], giving very different subsets as output. To overcome these problems we implement an appropriate experimental setup (adapted from a high-throughput molecular profiling methodology recently developed for microarray data [22]), consisting of two nested processes. The external process performs t times a random split of the data set in a learning set and a test set. The learning set is then used by the inner process to perform the selection of nested subsets of features (with any of the three methods described above) and to develop classifiers over these subsets. In this case we use two classifiers, RF and linear SVMs. Fig. 2 shows this internal process in some detail. The set of classifiers is then evaluated on the test set, which has not been used at any time during the selection and modeling. The outputs of the t replicated experiments are then used to perform the feature ranking and the estimation of its accuracy in a robust and unbiased way.

For both data sets, we replicate the feature selection process $t = 100$ times. In each replication, we split the data set at random

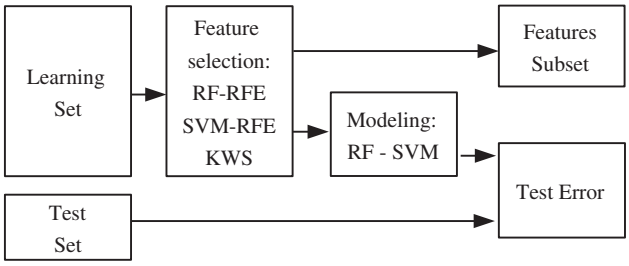

Fig. 2. Detail of the inner process of the experimental setup. This process is repeated $t=100$ times, starting from random splits in train and test sets. The output of the t replications is then used to estimate errors and to select important features.

into train/test sets with a 75%/25% proportion, keeping class frequencies balanced. Each one of the 100 train sets was used to select features with KWS, RF-RFE and SVM-RFE and to fit RF and SVM models over the selected subsets. The classifiers were then evaluated on the corresponding test sets using the usual mean classification error (ratio of incorrectly classified samples to total number of examples)

$$E_{\mathrm{T}} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}i), \tag{4}$$

where n is the number of samples in the test set, y_i is the true class, $\hat{y}_i$ is the class assigned by the classifier and $I(\ )$ is a function returning 1 if its argument is true and 0 otherwise.

3. Results and discussion

3.1. Modeling error

In Fig. 3 we compare the three selection methods on the Strawberry data set. In the top panel we show mean classification errors (±one standard deviation) for RF models developed over subsets selected with all methods. Results for all three methods show similar behaviors for more than 20

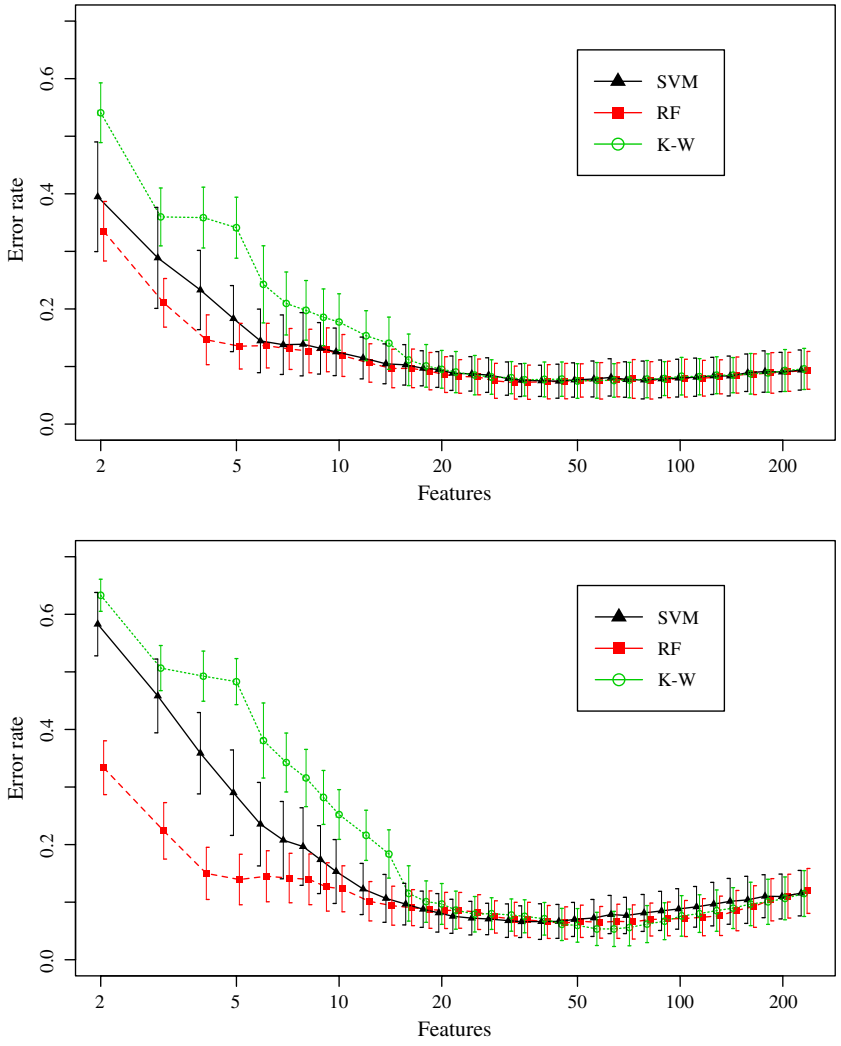

Fig. 3. Error levels for the three selection methods on the Strawberry data set. Top panel shows mean classification errors for Random Forest models. Bottom panel corresponds to Support Vector Machine models. Error lines show one standard deviation. On the legend, SVM stands for Support Vector Machine Recursive Feature Elimination, RF for Random Forest Recursive Feature Elimination and K–W for Kruskal–Wallis selections.

features, with also a similar minimum around 35 features. SVM-RFE has the minimum mean error, but the differences are very small in this case. However, for small subsets with less than 20 features there is a clear difference between the three, with RF-RFE giving the best performance followed by the other multivariate method, SVM-RFE. The increase in error levels when evaluating smaller subsets is typical of feature selection methods. After a given point in the feature selection loop all useless or redundant features have already been removed, and the algorithms begin to eliminate features that carry nonredundant information. An interesting question is whether the observed differences correspond to a better selection of features or just to a bias to choose features appropriate for RF, because the selection was made with the same algorithm. In the bottom panel of Fig. 3 we can see that SVM models (developed over the same subsets) give qualitatively similar results for less than 20 features. This confirms that RF-RFE selects small subsets of features with better discrimination capabilities than the other two methods. For more than 20 features the picture is somewhat different from the RF case. The minimum error for RF-RFE and KWS selections is reached with 60 features. SVM-RFE reaches a slightly worse minimum at around 40 features.

The same analysis was repeated for the Nostrani data set. In the top panel of Fig. 4 we compare the three selection methodologies with RF modeling. This data set is more difficult than the previous one, as can be seen from the higher mean errors and standard deviations. The differences between the three methods are smaller in this case. Results are similar for less than 8 features, but RF-RFE selections work better for subsets between 8 and 35 features. SVM-RFE and RF-RFE show a minimum prediction error at around 55 features, with SVM selections outperforming RF ones. KWS shows a worse minimum, also with a higher number of features. As in the Strawberry data set, RF results are confirmed by SVM modeling (bottom panel of Fig. 4). RF-RFE selection shows the best performance, both in terms of small subsets selection (for more than 5 features) as well as for minimum error.

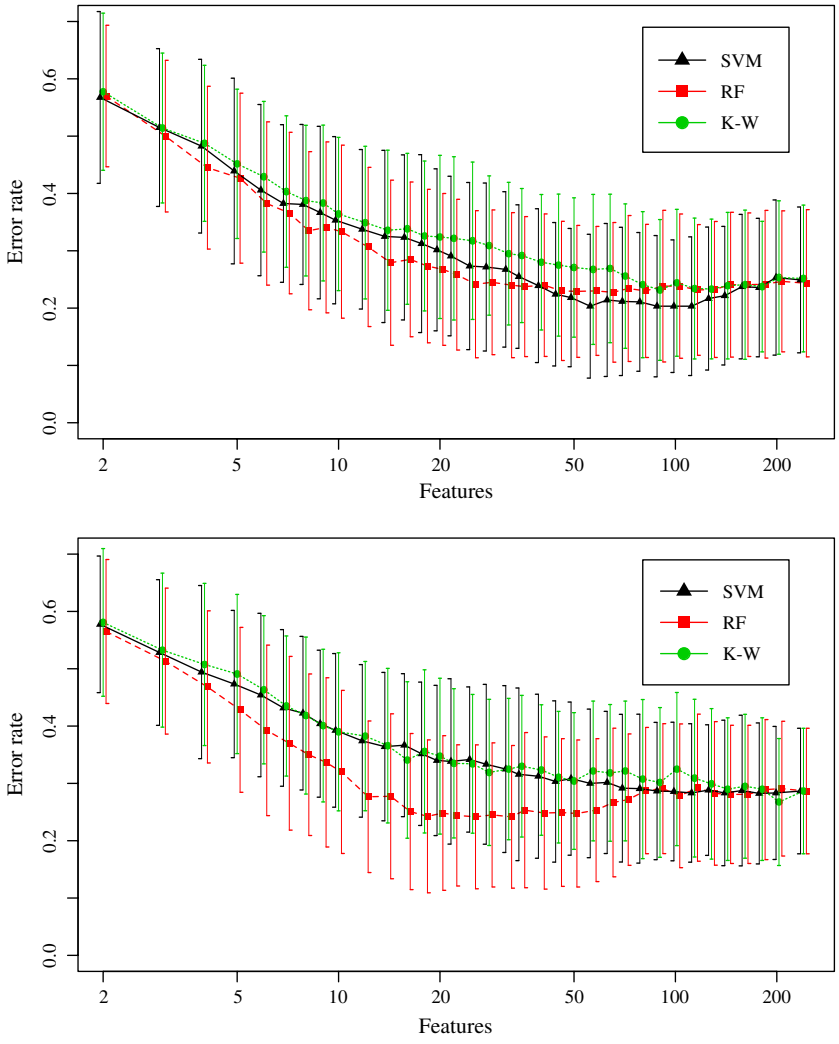

Fig. 4. Error levels for the three selection methods on the Nostrani data set. Top panel shows mean classification errors for Random Forest models. Bottom panel corresponds to Support Vector Machine models. Error lines show one standard deviation. On the legend, SVM stands for Support Vector Machine Recursive Feature Elimination, RF for Random Forest Recursive Feature Elimination and K–W for Kruskal–Wallis selections.

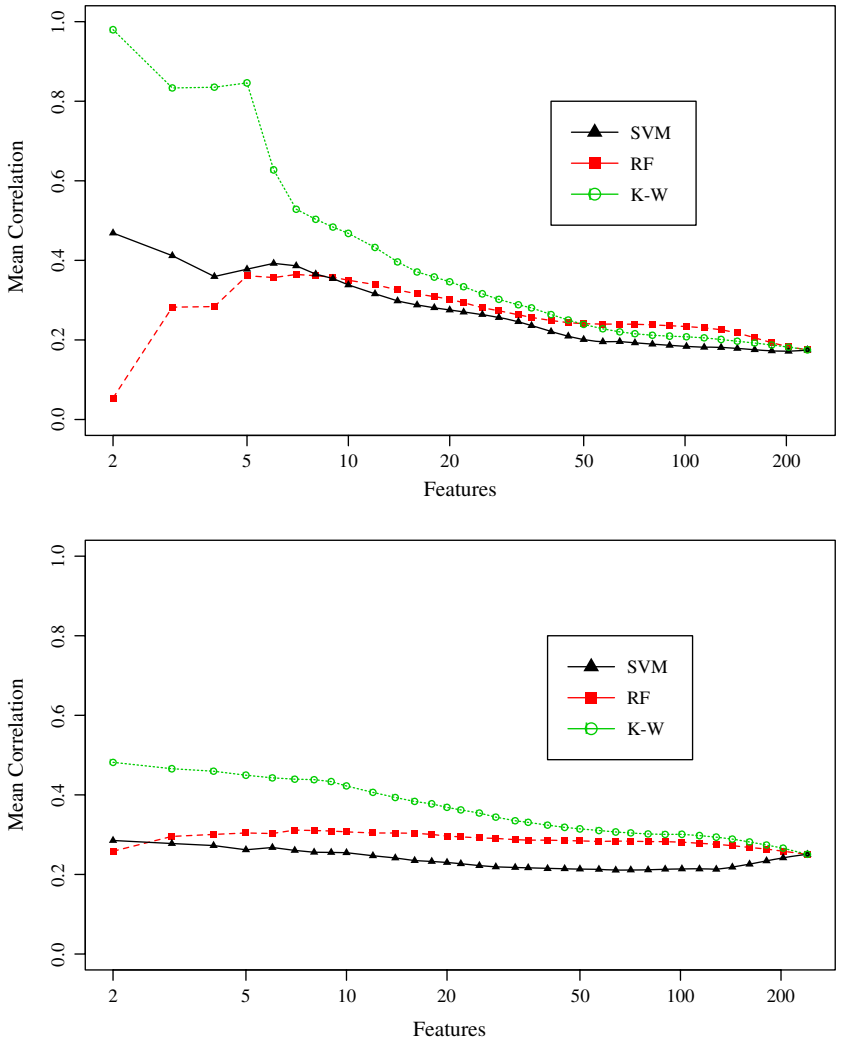

Fig. 5. Mean Correlation between subsets of different lengths (see text). Top panel: Strawberries data set. Bottom panel: Nostrani data set. On the legend, SVM stands for Support Vector Machine Recursive Feature Elimination, RF for Random Forest Recursive Feature Elimination and K–W for Kruskal–Wallis selections.

3.2. Feature correlations

The different performances of the selection methods, analyzed in the previous section, can be partially explained by the correlation between the selected features. For each subset selected by any of the three methods we calculate the correlation matrix[1] over the full data set, and take the mean value of the nondiagonal elements of that matrix. In Fig. 5 we show averaged values of this quantity over the 100 replicated experiments as a function of the number of features into the subsets. The top panel shows the mean correlation for the Strawberries data set. For more than 20 features the behavior is similar for the 3 methods, with only a small edge for SVM selections. On the left side of the picture we can see that KWS selects highly correlated features, opposite to RF-RFE that selects subsets with minimum correlation. The bottom panel shows mean correlation for the Nostrani data set. SVM selections show low correlations in this case. This is in agreement with the best modeling results showed by RF over SVM-RFE selections.

3.3. Finding relevant features

In this subsection we restrict our analysis to RF-RFE and SVM-RFE, which in all cases outperformed KWS. As stated in the Introduction, we need to relay on statistical procedures to analyze the relative importance of individual features, because selection methods are unstable and produce different rankings on each replicated experiment. For example, in the top panel of Fig. 3, there is a big difference in generalization error for subsets of 4 features selected by RF-RFE and SVM-RFE. In the top panels of Fig. 6 we show the selection probability associated with each member of these subsets of 4 features (estimated over the 100 replicated experiments). This probability is estimated by the fraction of times a given feature is selected into the subsets of a given size. In this case for RF selection (top-left panel) only 7 features have high selection probability. After the m/z ratio 57, there are 6 other m/z ratios with ~ 0.5 probability, which in fact

[1] A matrix with elements R_{ij} equal to the linear correlation coefficient between features i and j. The linear correlation coefficient is defined as $R = \sum x_k y_k$ for two series $\{x_k\}\{y_k\}$ with zero mean and unit variance.

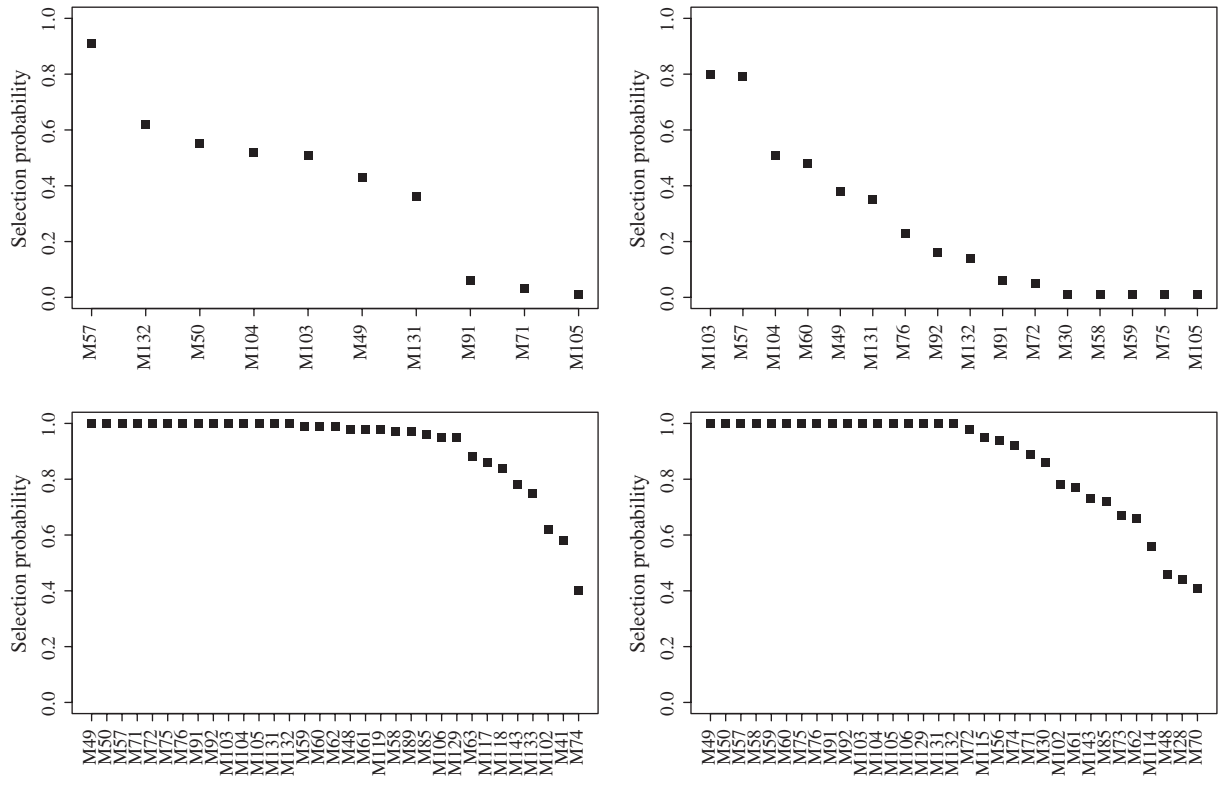

Fig. 6. Selection probability for different *m*/*z* ratios and subsets of different lengths, for the Strawberry data set. On the top line for 4 features subsets and on the bottom for 33 features subsets. On the left column for RF-RFE and on the right for SVM-RFE.

are 3 pairs (49–50, 103–104 and 131–132) of highly correlated features ($R>0.95$), corresponding to isotopes of the same compounds. More information can be extracted analyzing the co-occurrence of these features into the subsets. Table 1 shows the co-occurrence matrix for the Strawberry data set and subsets of 4 features. Each cell in the table shows the number of times both *m*/*z* ratios are selected together over the 100 replications. It is clear that one *m*/*z* ratio (and only one) from each pair is always into the 4 features subsets, showing a good discrimination between correlated *m*/*z* ratios for RF-RFE. These findings are also in agreement with the low correlation values for RF-RFE on the left of Fig. 5, top panel. For SVM selection (Fig. 6, top–right panel), *m*/*z* ratios 57 and 103 are almost always into the subsets. But *m*/*z* ratio 104 also has a ~0.5 probability, and it is selected together with 103 in ~42% of the times, which obviously reduced the model's discrimination capacity.

Table 1
Co-occurrence of *m*/*z* ratios into subsets of length 4 selected by RF-RFE

	M49	M50	M57	M103	M104	M131	M132
M49	–	0	38	19	25	15	28
M50	0	–	51	31	26	21	32
M57	38	51	–	47	45	32	57
M103	19	31	47	–	4	19	30
M104	25	26	45	4	–	18	32
M131	15	21	32	19	18	–	0
M132	28	32	57	30	32	0	–

The bottom panels of Fig. 6 show the 33 features with higher probabilities of being selected into 33 features subsets, corresponding to the minimum average error for both methods. Nearly 70% of the selected *m*/*z* ratios are the same. This explains the similar minimum generalization errors obtained with both methods. Also, for both methods there are at least 20 features with selection probability higher than 0.9, which can be used for the identification of relevant compounds.

For the Nostrani data set the differences between the methods are evident only for medium size subsets, with more than 20 features, which prevents a simple analysis of the correlations between *m*/*z* ratios, as presented for the Strawberries data sets. But similar results about features with high selection probabilities can be extracted in this case.

4. Conclusion

In this paper we have introduced the use of SVM-RFE and RF-RFE as efficient methods for feature selection for PTR-MS data. We used completely independent test sets coupled with replicated experiments in order to obtain unbiased and stable evaluations of their performance. Feature selection methods can be evaluated at least on two aspects, their capacity to find small subsets with a high discrimination capability, or to find the minimum possible error without caring about the number of selected features. On the first aspect, we showed that both methods outperform the use of univariate rankings and that RF-RFE has better performance than SVM-RFE on the two PTR-MS data sets under evaluation. Also, we have shown that these

differences in performance are independent of the classification algorithm, which points to a more efficient feature selection by RF-RFE for small data sets. The analysis of the mean linear correlation between the selected subsets supports this conclusion. On the second aspect, the search for the minimum possible error, we found no clear differences between the three methods. Only in one out of four cases (SVM models on the Nostrani data set) one of the methods outperforms the other two.

In all cases the minimum modeling error is reached by a classification method which was not used to select the features. This suggests that RFE methods in practice overfit the selection process, finding some particularities that are not repeated in the test sets. A different modeling strategy probably ignores those findings leading to a better generalization. More research about this combined strategy is currently underway.

We used selection probabilities and co-occurrences to highlight important m/z ratios. The analysis of the Strawberry data set indicates that the differences for very small subsets can be related with an improved discrimination between highly correlated features.

We will further investigate the effect of other strategies for multiclass extension or the use of nonlinear extensions in the SVM-RFE method. Ultimately, this highlights one of the main advantages of RF-RFE: it does not require any fine tuning to produce competitive results.

Acknowledgments

P.M.G. is supported by the PAT project SAMPPA. The PTR-MS analysis was partially supported by PAT projects MIROP and QUALIFRAPE.

References

[1] A. Hansel, A. Jordan, R. Holzinger, P. Prazeller, W. Vogel, Int. J. Mass Spectrom. Ion Process. 149/150 (1995) 609–619.
[2] W. Lindinger, A. Hansel, A. Jordan, Int. J. Mass. Spectrom. Ion Procs. 173 (1998) 191–241.
[3] F. Biasioli, F. Gasperi, E. Aprea, L. Colato, E. Boscaini, T.D. Märk, Int. J. Mass. Spectrom. 223–224 (2003) 343–353.
[4] F. Biasioli, F. Gasperi, E. Aprea, D. Mott, E. Boscaini, D. Mayr, T.D. Märk, J. Agric. Food Chem. 51 (2003) 7227–7233.
[5] E. Boscaini, S. Van Ruth, F. Biasioli, F. Gasperi, T.D. Märk, J. Agric. Food Chem. 51 (2003) 1782–1790.
[6] F. Boscaini, F. Gasperi, E. Aprea, I. Endrizzi, V. Framondino, F. Marini, D. Mott, T.D. Märk, Food Qual. Prefer. 17 (2006) 63–75.
[7] R. Kohavi, G.H. John, Artif. Intell. 97 (1996) 273–324.
[8] I. Guyon, A. Elissee, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[9] J.M. Sutter, J.H. Kalivas, Microchem. J. 47 (1993) 60–66.
[10] A. Alexandridis, P. Patrinos, H. Sarimveis, G. Tsekouras, Chemometr. Intell. Lab. Syst. 75 (2005) 149–162.
[11] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. 46 (2002) 389–422.
[12] S. Ramaswamy, et al., Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 15149–15154.
[13] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Chem. Inf. Comp. Sci. 44 (2004) 1630–1638.
[14] H. Li, C.Y. Ung, C.W. Yap, Y. Xue, Z.R. Li, Z.W. Cao, Y.Z. Chen, Chem. Res. Toxicol. 18 (2005) 1071–1080.
[15] P.M. Granitto, F. Gasperi, F. Biasioli, C. Furlanello, in: C.I. Chesñevar, J.C. Gomez (Eds.), Argentine Symposium on Artificial Intelligence 2005. 29–30 August 2005, Rosario, Argentina. Proceedings of the conference, (2005) 191–199.
[16] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[17] T.R. Golub, et al., Science 286 (1999) 531–537.
[18] K.J. Johnson, R.E. Synovec, Chemometr. Intell. Lab. Syst. 60 (2002) 225–237.
[19] W.H. Kruskal, W.A. Wallis, J. Am. Stat. Assoc. 47 (1952) 583621.
[20] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, Bioinformatics 21 (2005) 631–643.
[21] C. Ambroise, G. McLachlan, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 6562–6566.
[22] C. Furlanello, M. Serafini, S. Merler, G. Jurman, BMC Bioinformatics 4 (2003) 54.
[23] L. Breiman, Ann. Stat. 24 (1996) 2350–2383.
[24] F. Gasperi, F. Biasioli, V. Framondino, I. Endrizzi, Sci. Tec. Latt.-Casearia 55 (2004) 345–364.
[25] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[26] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.
[27] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[28] E. Allwein, R. Schapire, Y. Singer, J. Mach. Learn. Res. 1 (2000) 113–141.
[29] C.W. Hsu, C.J. Lin, IEEE Trans. Neural Netw. 13 (2002) 415–425.