Feature selection for classification of animal feed ingredients from near infrared microscopy spectra

José Antonio Sánchez del Rivero[a], Elena Montañés-Roces[a], Begoña de la Roza-Delgado[b], Ana Soldado[b], Oscar Luaces[a,*], José Ramón Quevedo[a], Antonio Bahamonde[a]

[a] *Artificial Intelligence Center*
University of Oviedo at Gijón, Asturias, Spain
[b] *Dept. of Nutrition, Grasslands and Forages*
SERIDA. Villaviciosa, Asturias, Spain

Abstract

The classification of animal feed ingredients has become a challenging computational task since the food crisis that arose in the European Union after the outbreak of Bovine Spongiform Encephalopathy (BSE). The most interesting alternative to replace visual observation under classical microscopy is based on the use of near infrared reflectance microscopy (NIRM). This technique collects spectral information from a set of microscopic particles of animal feeds. These spectra can be classified using maximum margin classifiers with good results. However, it is difficult to interpret the models in terms of the contribution of features. To gain insight into the interpretability of such classifications, we propose a method that learns accurate classifiers defined on a small set of narrow intervals of wavelengths. The proposed method is

*Corresponding author: Tel: +34 985 182 032
Email addresses: elena@aic.uniovi.es (Elena Montañés-Roces), broza@serida.org (Begoña de la Roza-Delgado), asoldado@serida.org (Ana Soldado), oluaces@aic.uniovi.es (Oscar Luaces), quevedo@aic.uniovi.es (José Ramón Quevedo), antonio@aic.uniovi.es (Antonio Bahamonde)

a greedy bipartite procedure that may be successfully compared with other state-of-the-art feature selectors and can be scaled up efficiently to deal with other classification tasks of higher dimensionality.

Keywords:

Feature selection, Interval selection, Variable selection, Spectroscopy, Near infrared, Ingredient discrimination

---

1. Introduction

The ban on the use of by-products of animal-origin such as meat and bone meal (*MBM*) in the feeding of farmed animals of the ruminant species [6] was one of the measures implemented in the European Union to stop the spread of bovine spongiform encephalopathy (BSE) and to prevent its reoccurrence. The official analytical method employed for the detection of banned ingredients in compound feedstuffs is classical microscopy [7]. This requires visual observation and interpretation by an experienced analyst. It is thus both tedious and subjective, so a number of different methods have been proposed to improve productivity and reduce costs.

In this paper we deal with datasets obtained from one of these alternatives, namely near infrared reflectance microscopy (*NIRM*) [12, 23, 26], which allows the collection of hundreds or thousands of spectra from a feed sample. NIRM has been proposed as a new analytical approach for identifying ingredients in animal feed and detecting undesirable substances such as MBM in feedstuffs.

NIRM is based on the collection of spectra from samples. These spectra can be collected from extremely small areas (<50 micrometers) using a

Fourier transform near infrared reflectance (*FT-NIR*) instrument attached to a microscope with an optical system designed to increase the efficiency of radiation transmission.

The spectra obtained from samples are used to determine the origin of feed ingredients in a sample. From a computational point of view, the spectra are numerical representations of substances, a vector of *absorbances* in intervals of wavelengths. Absorbance is a measure of the capacity of a substance to absorb light of a specified wavelength.

The classification of these spectra has been performed using different Machine Learning tools. Support Vector Machines (SVM) and other maximum margin learners have been shown to be valuable techniques for detecting banned MBM [23, 9]. However, the classifiers learned with these techniques are difficult to interpret. Although the classifications may be correct, the underlying causes may not be easily explained, a fact which decreases confidence in the results.

This paper focuses on improving the understandability of spectra classifications without losing accuracy. A popular approach employed for this purpose is to use Machine Learning tools that provide more explicit knowledge than hyperplanes, typically classification rules; but in this case, the accuracy could suffer. If we want to retain the accuracy of maximum margin learners, we may select a set of features. Then the learner will return a classifier built on a reduced set of wavelengths. This strategy to explain a classification procedure has been successfully employed in a number of application fields [4, 20].

Feature selection problems are NP-hard, so an exhaustive search is only

possible for datasets with a small number of features. For larger datasets we have to use approximate algorithms, which usually provide acceptable solutions at a reasonable computational cost. Sequential forward selection [31] and sequential backward selection [21] are the two basic approximate approaches. There are also metaheuristic approaches for feature selection such as genetic algorithms [34, 14], and methods based on rough set theory [33, 3, 15, 2, 35] and on Boolean independent component analysis [1].

However, the selection method has to be carefully chosen. Since the set of features itself will be part of the solution to the classification task, we shall present a method to produce accurate classifiers defined on a small set of narrow intervals of wavelengths. In fact, the biochemical meaning of spectra is attached to intervals of wavelengths that are related to the molecular structure of the ingredients.

The core idea is a feature selection process that prefers intervals instead of disconnected subsets of features. The proximity of the features in terms of their wavelength values must be taken into account. Tibshirani et al. [27] designed a generalization of lasso for regression problems whose features are ordered in a meaningful way. Their approach, named 'fused lasso', is biased towards sparse solutions and local constancy in the coefficients profile.

There are many application fields in which the contiguous nature of features plays an important role. This is the case, for instance, of speech, written texts, and music. In addition, genomic information is also arranged linearly. In this field, Kim and Xing [16] proposed a method to identify a small subset of contiguous blocks of single-nucleotide polymorphisms (SNPs) associated with a given phenotype. As in [27], the work reported in [16] is devised to

solve regression problems by forcing the sparsity of the solution. However, in this case a Laplacian prior is placed on the regression coefficients, instead of the L_1 penalty used in the fused lasso. Furthermore, both approaches were designed to handle situations with more features than training examples, which is the typical setting in genetic datasets.

A different approach can be found in paper by Leardi and Nørgaard [18]. Here, the authors propose to obtain a predictive model by first selecting intervals of features by means of backward interval partial least squares (biPLS) and then a genetic algorithms (GAs) applied on the resulting subset of features.

The idea of using a two-stage strategy is somewhat similar to the one presented in this paper. However, our approach is based on making a recursive bipartition of the input space to check whether any of the halves can be discarded by a learner. In fact, the proposed method is a wrapper which applies this bipartition greedily. It has a reasonably low time complexity in order to be scalable and hence useful in high dimensional input spaces. As stated previously, we have applied our method to classification tasks, although its adaptation to tackle regression problems is straightforward.

The paper is organized as follows. After presenting the method, we report a set of experiments carried out to test its benefits. We compare our greedy method with five other feature selection approaches: a *Backward* (*Back*) selector, a *Forward* (*Fwd*) selector, a selector built on the Recursive Feature Elimination (*RFE*) ranker [13], an interval-oriented *Relief* [17] selector, and the iPLS method proposed by Leardi and Nørgaard [18]. The results show that our method yields intervals which allow the learner to be as accurate as

when classical selectors are used, but it is much faster and produces slightly less intervals with fewer wavelengths. In other words, the classifiers produced by the greedy method are scalable and they better explain the underlying reasons used to discriminate animal feed ingredients.

2. Near Infrared Spectra

A spectral library was built using the most common ingredients included in feedstuffs together with banned ingredients such as processed animal proteins. The samples were provided by the major feed industries and rendering plants in the north of Spain from 2005 to 2009, thus representing the variability encountered in the actual production process. All the ingredients included in this study and the numbers of spectra collected are listed in Table 1. The optimal methodology for sample pre-treatment and the instrumental conditions to collect spectral data in this system were previously investigated by del Valle Fernández-Ibáñez et al. [28]. In line with their findings, the samples were ground to a particle size of 1 mm as the sole pre-treatment prior to NIRM analysis.

The datasets used in this paper were collected using an Auto Image Microscope connected to a PerkinElmer Spectrum One Fourier Transform Near Infrared (FT-NIR) Spectrometer in reflectance mode. The spectra were measured using fields of view of 50×50 micrometers arranged in a 13×18 grid over this area, collecting approximately 200 spectra per sample. This method avoids any subjective selection.

The spectra were obtained from the ratio between raw spectra and the background. The spectral information was stored as $\log(1/R)$, where R is

the *reflectance*, recorded at 4 nm intervals over the range 1112–2492 nm after conversion from cm^{-1} using the PerkinElmer software, Spectrum v. 5.01. We thus have $1 + (2492 - 1112)/4 = 346$ features per spectrum.

Figure 1 shows the spectra of 3 different animal feed ingredients of: wheat, soybean meal, and meat and bone meal (MBM).

3. Greedy Bipartition Methods

In this paper we deal with a collection of binary classification tasks, each consisting of separating spectra from a couple of groups of ingredients. Before presenting the feature selection tools, let us start by formally defining a binary classification task in addition to giving a brief description of the base learner used, a regularized logistic regressor, *LibLinear* [19, 8].

Formally, a binary classification task is represented as a set $D = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, l\}$, where inputs $\boldsymbol{x}_i \in \mathbb{R}^n$ are real vectors of dimension n representing spectra, and $y_i \in \{1, -1\}$ stand for the classes.

From the dataset D, *LibLinear* induces a probability model

$$\Pr(y = \pm 1|\boldsymbol{x}) = \frac{1}{1 + e^{-y(\boldsymbol{w}^T \boldsymbol{x} + b)}}, \tag{1}$$

where $\boldsymbol{w}$ and b are learning parameters. The classifier learned is then given by

$$\text{sign}(\Pr(class = +1|\boldsymbol{x}) - 0.5). \tag{2}$$

The parameters $\boldsymbol{w} \in \mathbb{R}^n$, and $b \in \mathbb{R}$, are *learned* by minimizing the negative log-likelihood

$$\min_{\boldsymbol{w}, b} \sum_{i=1}^{l} \log\left(1 + e^{-y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)}\right). \tag{3}$$

To obtain good generalization abilities, the authors of *LibLinear* added a regularization term, $\frac{1}{2}[\boldsymbol{w}; b]^T[\boldsymbol{w}; b]$, used in the formulation of SVM to incorporate the *maximum margin* principle. *LibLinear* thus solves the following convex optimization problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}[\boldsymbol{w}; b]^T[\boldsymbol{w}; b] + C\sum_{i=1}^{l} \log\left(1+e^{-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)}\right). \tag{4}$$

The value of the regularization parameter, C, is decided by users so that both terms in (4) are balanced.

Using *LibLinear* as the base learner, greedy bipartite methods proceed as follows. The range of wavelengths is split into two parts of equal size, *left* and *right*. An estimation of the accuracy with and without each part then decides if it is possible to get rid of one of the parts. The method proceeds recursively with each retained part until the size of the intervals falls below a threshold, ϵ. A description of the method is detailed in Algorithm 1. A sketch of the procedure is shown in Figure 2.

Notice that the roles of *left* and *right* may be changed. The method that start from left to right will be called *Left-Right (LR)*, while *Right-Left (RL)* will stand for the method that proceeds in the opposite way.

The implementation of the algorithm requires the specification of a procedure for searching for the best regularization parameter C, and a way to estimate the accuracy achieved with a subset of features, in the algorithm Test In. We employed a simple hold-out method in the experiments reported in the next section. Training sets were split into proper training sets plus a validation set for searching for the best value for C and to estimate the accuracy of a subset of features.

Finally, the algorithm returns the best set of intervals found together with a predictive model learned using only the selected features and the best estimation for C. This model is, of course, obtained from the whole training set, combining the corresponding training and validation sets.

4. Experimental Results

In this section we report an experimental comparison of the previously described methods. A total of 217 samples from 23 ingredients were analyzed. These ingredients were categorized in 7 groups. About 200 spectra were collected from each sample, giving a total of 46278 spectra. Each spectrum is described by 346 features, the values of absorbance for wavelengths in the interval $[1112, 2492]$ nm obtained each 4 nm, and belongs to one of the 7 groups of ingredients, which will be the target class to be predicted. We built 21 binary classification tasks with these data (combinations of 7 elements taken two by two).

All spectra were smoothed out using a moving average filter applied 10 times. The value of each original feature was thus transformed using the following function

$$f_i^j = \begin{cases} \dfrac{2f_i^{j-1} + f_{i+1}^{j-1}}{3} & \text{if } i = 1 \text{ (first)} \quad (5a) \\ \dfrac{f_{i-1}^{j-1} + 2f_i^{j-1}}{3} & \text{if } i = 346 \text{ (last)} \quad (5b) \\ \dfrac{f_{i-1}^{j-1} + 2f_i^{j-1} + f_{i+1}^{j-1}}{4} & \text{otherwise,} \quad (5c) \end{cases}$$

where $i = 1, \ldots, 346$, $j = 1, \ldots, 10$ and f_i^j is the i-th feature after j applications of the smoothing filter.

The greedy bipartite methods presented in Section 3 were implemented using $\epsilon = 10$; i.e., only intervals containing more than 10 wavelengths will be split. To test the convenience of skipping an interval of features in Algorithm (1), we separated training sets into validation (40% of the input examples) and proper training (the remaining 60%). This split was also used to tune the regularization parameter, C, with an internal grid search in $\{0.01, 1, 100\}$.

We compared the *LR* and *RL* versions or our approach with several standard selection procedures, as well as with a state-of-the-art native interval selector.

The standard procedures include a *Backward* (*Back*) and a *Forward* (*Fwd*) approach, which have been previously used in the field of chemometrics as greedy subset selectors [5, 24, 25]; a Recursive Feature Elimination (*RFE*) ranker proposed by Guyon et al. [13], which has proven very effective for sorting attributes according to their relevance in classification tasks; and, finally, the well-known *Relief* algorithm proposed by Kira and Rendell [17]. These selectors were adapted for interval selection since they were originally conceived to deal with individual features. To have comparable selections of intervals, the set of 346 features was packed into 34 intervals of 10 contiguous wavelengths and a final interval including the last 6 features to allow these methods to remove intervals of features. The final step of these approaches consists in joining contiguous intervals in order for them to be considered as a single wider interval.

The *Back* method starts with all attributes and searches for the interval that yields the highest estimated accuracy when discarded. That interval

is removed and the process is repeated with the remaining intervals until the method comes across a subset that cannot be reduced without a loss in accuracy. Conversely, the *Fwd* method starts with no features and progressively joins intervals until the estimated accuracy stops increasing. The interval added in each step is the one that yields the subset with the highest estimated accuracy.

The *RFE* and the *Relief* selectors start by constructing a ranking of intervals sorted by relevance. They then select the subset of top ranked intervals yielding the best estimated accuracy. The difference between the two approaches lies in the way they compute the ranking of intervals.

The general *RFE* procedure to build the ranking proceeds as follows

a) *RFE* constructs a predictive model, like the one in (1),

b) It removes the feature whose squared coefficient, w_i^2, is the lowest,

c) It applies *RFE* to the remaining features until all but one are removed.

We used an alternative criterion for step b) to allow the algorithm to consider intervals of features instead of individual features. Thus, *RFE* will remove the k-th interval of features whose sum of squared coefficients is the lowest, i.e.,

$$\arg\min_{n=1,\ldots,|I|} \sum_{i \in I_n} w_i^2$$

where I is the set of 35 intervals of features that make up the datasets.

We used the implementation of *RFE* provided in The Spider [29] toolbox, which supports this removing criterion for groups of variables, although it is not documented in the paper by Guyon et al. [13].

The adaptation made to *Relief* in order to select intervals instead of single features is similar to the one made for *RFE*. The score for each interval of

variables is thus computed as the sum of scores of their variables. In turn, the score for each variable is computed using the classical *Relief* as follows. For each training example, the algorithm finds the k nearest neighbors of its same class (*hit* examples) and the k nearest neighbors of the opposite class (*miss* examples) and computes the corresponding distance vectors. All distance vectors to the hit examples are averaged, as well as the distance vectors to the miss examples. Then, the score for each input variable x_i is the ratio between the average distance to the miss examples and the average distance to the hit examples projected on each variable x_i.

Furthermore, we also included a native interval selector, the *biPLS* procedure proposed by Leardi and Nørgaard [18]. The implementation does not allow the user to configure the intervals, but allows their minimum size to be defined, so we applied the algorithm guaranteeing that the intervals had at least 10 features. This partition practically yielded the same intervals used by the previously cited standard feature selectors.

In addition to the aforementioned selectors, we also considered a null selector, i.e., a learner that used all available features, *All*.

4.1. Estimation of performance by cross-validation

The estimation of accuracy was carried out using a 4-fold cross-validation procedure. We thus made four random splits, each one taking 75% of the data for training. As we have already explained, the training data is in turn split into two blocks (60%/40%) for parameter tuning purposes; see Figure 3. Once a combination of parameters is selected, we obtain a model using all the training data (the 75% mentioned above) and the resulting classification model is applied to the remaining 25% reserved for testing. The scores are

reported in Table 2. Note that the spectra of any sample were never split in different folds. The idea is to avoid spectra from the same sample split appearing in both the training and test set. Moreover, the folds built for this purpose had a balanced distribution of classes similar to that of the original classification task. This is the reason for using a cross-validation procedure with only 4 folds.

Although the computational complexity is not of core importance in this case, it is useful to analyze it in order to evaluate the scalability of the methods. Table 3 accordingly shows the running time in seconds needed to obtain a classification model after the feature selection on each dataset. The times were taken on a dedicated computer to avoid delays due to other users' processes. We also guaranteed enough memory for the experiments to prevent delays due to memory swapping to disk.

Obviously, the fastest method is the one that makes no selection at all. It was included just to show how long it takes to learn a model for each problem in the comparison. Among the techniques that make some feature selection, our greedy methods are the fastest in all but 4 problems, where they ranked in second position. Their average times were below 225 seconds. *RFE* and *Relief* have higher average times than our approaches, taking up to more than 350 seconds in the case of *Relief*.

The times taken for the remaining methods confirm that none of them are scalable solutions, so they can hardly be applied to datasets of moderate or large size. The case of *biPLS* is noteworthy on account of being extremely slow. We wish to emphasize that we did not implement this method; instead, we used the original source code provided by the authors, so the poor

performance cannot be imputed to incorrect implementation of the method.

However, the most important measures to evaluate the *understandability* of the hypotheses learned are those that deal with the number and distribution of features selected. Table 4 shows the number of discarded features and the number of intervals in which the selected features are grouped. These scores were computed by each method when they are applied to the whole dataset.

With the aim of summarizing in a single number how explicative a set of selected features is, we define the *quality* of the selection as follows

$$\begin{aligned} quality &= \frac{\#features\ excluded}{\#intervals} \\ &= \frac{346 - \#features\ selected}{\#intervals}. \end{aligned} \tag{6}$$

We chose this measure because the quality of the selection must be proportional to the number of features excluded and inversely proportional to the amount of intervals selected. Thus, for a given number of features excluded by two selection procedures, we will prefer the selection with a lower number of intervals. In turn, for a given number of intervals, we will prefer the selection method which yields smaller intervals, i.e., the one with a higher number of features excluded. The scores for this measure are also shown in Table 4.

4.2. Discussion

Following the recommendations of [11], we performed a two-step comparison for each of the considered measures: a Friedman test followed by a *post-hoc* pairwise comparison, namely a Bergmann-Hommel procedure. This

is a non-parametric test that starts computing the average ranking positions of each method across the datasets considered.

Table 5 shows the average ranking positions for the different selection approaches. The best approach for each performance measure is the one with the average ranking position closest to 1 and is highlighted in bold.

Considering a level of significance $\alpha = 0.05$, the difference in accuracy is statistically significant between the worst (*biPLS*) and the rest of the approaches except for *Fwd*, which is the last but one in the ranking of accuracy. If we raise this level up to $\alpha = 0.1$, then the difference between *Back* and *Fwd* is also significant. We have also included in the comparison the results obtained with no selection method, i.e. when using *All* features, which occupies the second position in the accuracy ranking.

As can be seen in Table 2, one of the most successful results is the discrimination between cereals, which are the most usual ingredients in animal feeds, and banned ingredients (1 vs 7). The accuracy obtained in the distinction between groups 4 and 7 is also worth noting. In fact, the most important confusion in classification procedures previously reported was between soybean meal (included in group 4), of vegetable origin, and MBM (included in group 7), of animal origin [10]. Confusion may arise due to the high protein levels in both ingredients. On the other hand, the accuracy decreases between vegetable ingredients (2 vs 6), which could be due to the fact that they have similar cell wall structures.

In terms of quality, our proposed approaches, *LR* and *RL*, are top ranked, closely followed by *Fwd*. The differences in terms of quality between any of these three approaches and the remaining selection methods are statistically

significant for both $\alpha = 0.05$ and $\alpha = 0.1$. The quality measure proposed in (6) depends on the number of features discarded by an algorithm and the number of intervals selected, so we also analyzed these measures separately.

The *Fwd* approach is the one that discards most features, though the difference is only significant with respect to the worst approaches, *Back*, *RFE* and *Relief*. Thus, our approaches are not significantly worse than the best, *Fwd*, when discarding features. Moreover, *LR* and *RL* are also significantly better than *Back*, *RFE* and *Relief* at $\alpha = 0.05$.

Furthermore, we are not only interested in discarding as many features as possible, but also in selecting few and small (if possible) intervals or regions of the spectrum in order to gain insight into the process of discriminating between ingredients. As regards the number of intervals selected, *biPLS* is the worst approach, followed by *Fwd*. Note, however, that *Fwd* is the approach which, on average, discards the largest number of features. This means that, in general, *Fwd* selects less features than the other approaches, but its selection is more fragmented over the spectrum. In terms of number of intervals, *Relief* is ranked best in the comparison, closely followed by our greedy bipartite methods. There are no significant differences among them and they select a significantly lower number of intervals than *Fwd* and *biPLS* at $\alpha = 0.05$.

Figure 4(a) depicts an example of interval selection by our greedy bipartite methods in the problem of discriminating between by-products (class 5) and banned ingredients (class 7). It also shows a random sample of 5 spectra of each class taken for the training data.

The main characteristic bands differentiating the two classes of spectra

(permitted and banned) are clearly visible, and there are considerable differences in the location and shape of the bands, in the region characteristic of fat absorption (1724–1760 nm). According to Murray et al. [22], this band is related to the content in polyunsaturated fatty acids. On the other hand, in [32] protein absorption was related to band (2054-2274 nm). In the field of animal nutrition, it is well known that most of the banned ingredients are derived from terrestrial animal tissues and they have relatively high fat and protein contents. This means that differences occur at specific wavelength ranges. Carbohydrates in general may have a free OH stretch absorption near 1440 nm [30]; in this regard, Williams [32] reported an important band correlated with cellulose carbohydrate (fiber content) 1490 nm.

In this application, the classifications of spectra may be similar using both strategies, as the principal intervals selected when processing the data from the left to right of the spectral data and vice versa are mainly related to protein content.

In this particular example, and contrary to its usual behavior, *Fwd* discards only a few features, giving only 2 large intervals which are not very informative, since they cover almost the entire spectrum. In turn, the *RFE* procedure selects 6 intervals, one of which is rather broad, so it is also barely informative. These results are depicted in Figures 4(c) and 4(b), respectively. The *Back* approach also selects 6 intervals, although they are more compact than those selected by *RFE*, as can be seen in Figure 4(d). The *Relief* approach made a similar selection to that made by *RFE*, except for the band in the lowest wavelengths, which was discarded by *Relief*. Finally, *biPLS* selected 7 wide intervals, discarding few features, thus being as uninformative

as *Fwd* in this particular example.

Finally, we also compared the selection approaches with respect to their running time. Our approaches significantly outperform all the other selectors at $\alpha = 0.1$, although the difference with *RFE* is not significant at $\alpha = 0.05$. In fact, differences are very important in most of the absolute scores (Table 3) and hence in the average ranking positions (Table 5).

5. Conclusion

The classification of spectra of animal feed ingredients has to be accurate, but it also should be understandable. In this context, the chemical or biological explanation is related to the identification of a reduced set of contiguous features grouped in few intervals (spectral bands).

On the other hand, it has been shown that the combination of NIRM spectroscopy and a maximum margin classifier should allow a regulatory laboratory to certify and quantify the presence of meat and bone meal in common samples of processed animal feed [23, 9]. However, the classifiers obtained are difficult to understand.

In this paper we have presented some methods to make the classifiers learned by maximum margin classifiers from binary classification tasks more explicative. Using this criterion, the greedy bipartite methods (*LR, RL*) significantly outperform the other feature selection strategies compared in the paper in terms of quality. They are closely followed by *Fwd*, though this method is much worse in terms of computational complexity. Furthermore, neither *LR* nor *RL* presents significant differences in accuracy with respect to the most accurate method found in the comparison. In addition, *LR* and *RL*

identify few and very compact intervals of features. Thus, the classifiers obtained by LR and RL enable the explanation of accurate classifications when dealing with animal feed ingredients. Moreover, this methodology opens up other alternative methods for checking feed composition.

Finally, we have shown that implementing a feature selector for interval selection is not a straightforward task, especially if the aim is to produce scalable methods able to deal, eventually, with datasets of very high dimensionality.

Acknowledgments

We would like to thank R. Leardi and L. Nørgaard for providing us with the source code of their algorithm, *biPLS*, to be included in our experimental comparison.

The research work by the Artificial Intelligence Center reported here is supported in part under grant TIN2011-23558 from the Spanish Ministerio de Economía y Competitividad. The SERIDA work was supported by the Spanish project RTA2010-00128-00-00 from the INIA.

[1] B. Apollonia, S. Bassisa, A. Brega, Feature selection via boolean independent component analysis, Information Sciences 179 (2009) 3815–3831.

[2] D. Chen, C. Wang, Q. Hu, A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets, Information Sciences 17 (2007) 3500–3518.

[3] C. Cornelis, R. Jensen, G. Hurtado, D. Ślezak, Attribute selection with fuzzy decision reducts, Information Sciences 180 (2010) 209–224.

[4] J.J. del Coz, G.F. Bayón, J. Díez, O. Luaces, A. Bahamonde, C. Sañudo, Trait selection for assessing beef meat quality using non-linear SVM, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17 (NIPS '04), MIT Press, Cambridge, MA, 2005, pp. 321–328.

[5] F. Du, Y.J. Li, T.J. Wu, Regularized orthogonal forward feature selection for spectral data, in: Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on, pp. 645 –650.

[6] European Commission, Regulation, EC 999/2001 of the European Parliament and of the Council laying down rules for the prevention, control and eradication of certain transmissible spongiform encephalopathies, Official Journal of the European Communities 147 (2001) 1–40.

[7] European Commission, Commission Directive 2003/126/EC of 23 December 2003 on the analytical method for the determination of constituents of animal origin for the official control of feedingstuffs, Official Journal of the European Communities 339 (2003) 78–84.

[8] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9 (2008) 1871–1874.

[9] V. Fernández-Ibáñez, T. Fearn, E. Montañés, J.R. Quevedo, A. Soldado, B. de la Roza-Delgado, Improving the Discriminatory Power of a Near-

Infrared Microscopy Spectral Library with a Support Vector Machine Classifier, Applied Spectroscopy 64 (2010) 66–72.

[10] V. Fernández-Ibáñez, T. Fearn, A. Soldado, B. de la Roza-Delgado, Development and validation of near infrared microscopy spectral libraries of ingredients in animal feed as a first step to adopting traceability and authenticity as guarantors of food safety, Food Chemistry 121 (2010) 871–877.

[11] S. García, F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.

[12] G. Gizzi, L.W.D. Van Raamsdonk, V. Baeten, I. Murray, G. Berben, G. Brambilla, C. Von Holst, An overview of tests for animal tissues in feeds applied in response to public health concerns regarding bovine spongiform encephalopathy, Revue scientifique et technique-Office international des épizooties 22 (2003) 311–331.

[13] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.

[14] W. Hsu, Genetic wrappers for feature selection in decision tree induction and variable ordering in bayesian network structure learning, Information Sciences 163 (2004) 103–122.

[15] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heteroge-

neous feature subset selection, Information Sciences 178 (2008) 3577–3594.

[16] S. Kim, E. Xing, Feature selection via block-regularized regression, in: Proceedings of the 24th Conference on Uncertainty in AI (UAI).

[17] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Conference on Machine Learning, Morgan Kaufmann, 1992, pp. 249–256.

[18] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, Journal of Chemometrics 18 (2004) 486–497.

[19] C.J. Lin, R.C. Weng, S.S. Keerthi, Trust Region Newton Method for Logistic Regression, Journal of Machine Learning Research 9 (2008) 627–650.

[20] O. Luaces, G.F. Bayón, J.R. Quevedo, J. Díez, J.J. del Coz, A. Bahamonde, Analyzing sensory data using non-linear preference learning with feature subset selection, in: J.F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD '04), Pisa, Italy, pp. 286–297.

[21] S. Maldonado, R. Weber, A wrapper method for feature selection using support vector machines, Information Sciences 179 (2009) 2208–2217.

[22] I. Murray, L.S. Aucott, I.H. Pike, Use of discriminant analysis on visible and near infrared reflectance spectra to detect adulteration of fish-

meal with meat and bone meal, Journal of Near Infrared Spectroscopy 9 (2001).

[23] J.A. Pierna, V. Baeten, A.M. Renier, R.P. Cogdill, P. Dardenne, Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds, Journal of Chemometrics 18 (2004) 341–349.

[24] F. Rossi, D. Francois, V. Wertz, M. Meurens, M. Verleysen, Fast selection of spectral variables with b-spline compression, Chemometrics and Intelligent Laboratory Systems 86 (2007) 208 – 218.

[25] F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems 80 (2006) 215 – 226.

[26] B. de la Roza-Delgado, A. Soldado, A. Martínez-Fernández, F. Vicente, A. Garrido-Varo, D. Pérez-Marín, M.J. de la Haba, J.E. Guerrero-Ginel, Application of near-infrared microscopy (NIRM) for the detection of meat and bone meals in animal feeds: A tool for food and feed safety, Food Chemistry 105 (2007) 1164–1170.

[27] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of Royal Statistical Society Series B 67 (2005) 91–108.

[28] M. del Valle Fernández-Ibáñez, A. Soldado, F. Vicente, A. Martínez-Fernández, B. de la Roza-Delgado, Particle size optimisation in devel-

opment of near infrared microscopy methodology to build spectral libraries of animal feeds, Journal of Near Infrared Spectroscopy 16 (2008) 243–248.

[29] J. Weston, A. Elisseeff, G. BakIr, F. Sinz, Spider: object-orientated machine learning library, 2005.

[30] L. Weyer, S.C. Lo, Spectra-structure correlations in the near-infrared, in: Handbook of Vibrational Spectroscopy, volume 3, Wiley, U.K., 2001, pp. 1817–1837.

[31] A. Whitney, A direct method of nonparametric measurement selection, IEEE Transactions on Computers C-20 (1971) 1100–1103.

[32] P. Williams, Near infrared technology in the agricultural and food industries, Near infrared technology in the agricultural and food industries, AACC Press, 2001, pp. 145–169.

[33] Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Information Sciences 178 (2008) 3356–3373.

[34] M. Zhang, J. Peña, V. Robles, Feature selection for multi-label naive Bayes classification, Information Sciences 179 (2009) 3218–3229.

[35] S. Zhao, E. Tsang, On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences 178 (2007) 3163–3176.

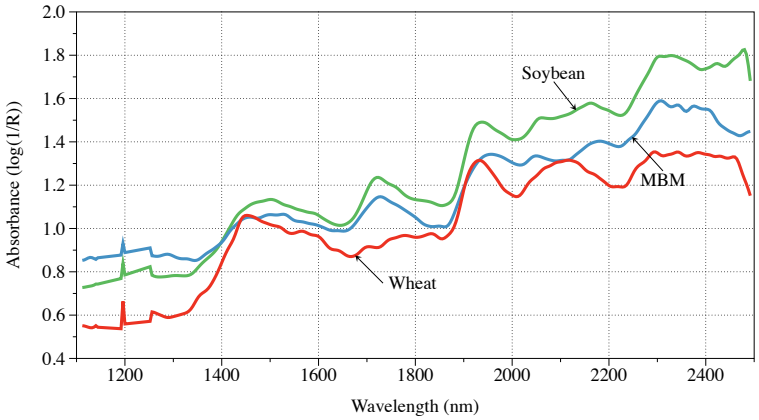


Figure 1: Spectra of 3 different types of animal feed ingredients: *wheat*, *soybean* meal, and *MBM* (meat and bone meal).

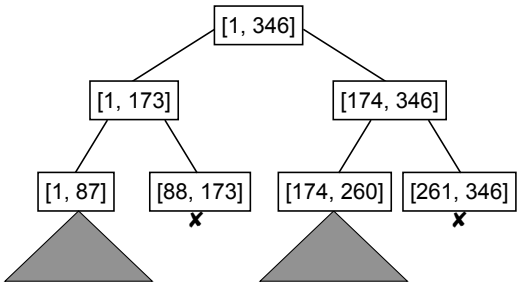


Figure 2: Search tree used by one of the greedy bipartite methods for feature selection.

Algorithm 1 Greedy bipartition algorithm to select a set of intervals of features. As it starts testing left and then right parts, it is the LR version. Changing the roles of left and right, we obtain the RL version.

```
Input: Interval
Open ← [Interval]; {list of intervals}
best.interval ← Interval;
best.accuracy ← Test In(Interval);
repeat
  Interval ← First(Open);
  Open ← Rest(Open);
  (left, right) ← Bipartite (Interval);
  but_left ← Test In(best.interval - left);
  if (but_left > best.accuracy) then
    best.interval ← best.interval - left;
    best.accuracy ← but_left;
    Open ← Open + [right];
  else
    Open ← Open + [left];
    but_right = Test In(best.interval - right);
    if (but_right > best.accuracy) then
      best.interval ← best.interval - right;
      best.accuracy ← but_right;
    else
      Open ← Open + [right];
    end if
  end if
until (length(left) < ε or length(right) < ε)
return best.interval;
```

Group	Ingredients	Samples	Spectra
Group 1	Oats	4	847
Cereals	Rye	5	1146
	Barley	20	4216
	Wheat	15	3210
	Maize	21	4826
	Total	65	14245
Group 2	Dehydrated lucerne	24	5109
Forages	Cereal straw	27	5644
	Grass hay	2	421
	Fababean silage	2	450
	Grass silage	13	2713
	Total	68	14337
Group 3	Cotton seed	7	1016
Fat concent.	Sunflower seed	7	1578
	Total	14	2594
Group 4	Peas	2	427
Protein concent.	Soybean meal	14	3120
	Total	16	3547
Group 5	Corn flakes	1	210
By-products	DDGS Barley	1	202
	Bran	2	419
	Beet pulp	14	3039
	Total	18	3870
Group 6	Maize silage	23	4836
Maize silage	Total	23	4836
Group 7	Meat and bone meal	10	2146
Banned	Blood meal	1	235
ingredients	Hemoglobin	1	234
	Animal plasma	1	234
	Total	13	2849

Table 1: Groups of animal feed ingredients. For each group, we report the number of samples and spectra.

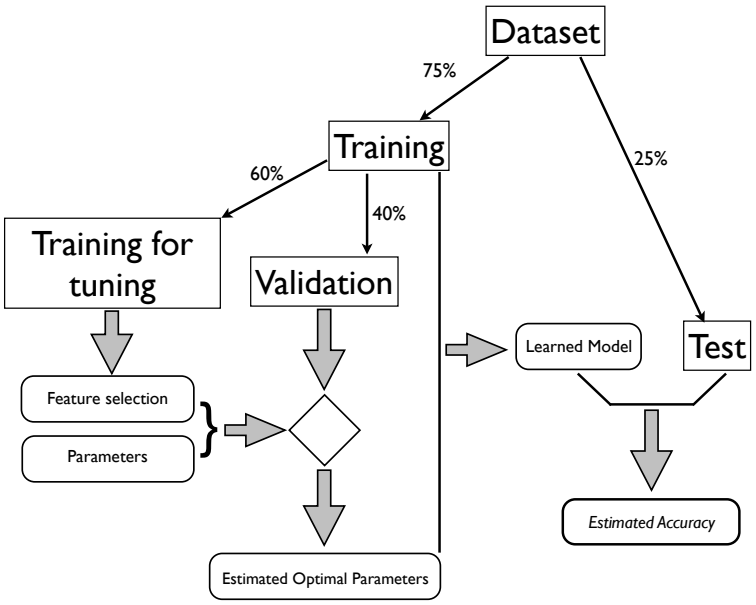


Figure 3: Graphical representation of the accuracy estimation for each fold in the cross-validation.

Dataset	LR	RL	Back	Fwd	RFE	Relief	biPLS	All
1 vs 2	98.06	**98.09**	98.01	97.84	97.97	97.91	97.09	98.07
1 vs 3	97.56	98.63	**98.80**	98.10	98.00	97.83	96.74	98.67
1 vs 4	96.87	96.95	96.88	96.80	**96.95**	96.86	96.75	96.90
1 vs 5	95.33	95.36	**95.45**	95.18	95.37	95.28	94.08	95.30
1 vs 6	96.80	95.70	96.90	**97.49**	96.51	96.47	95.48	95.89
1 vs 7	99.86	99.75	99.92	**99.97**	99.94	99.91	99.79	99.94
2 vs 3	94.49	94.40	**94.53**	93.95	94.51	94.47	91.99	94.50
2 vs 4	98.53	98.69	98.58	98.64	**98.73**	98.60	97.33	98.62
2 vs 5	93.68	93.47	93.67	93.27	93.71	**93.84**	91.10	93.74
2 vs 6	87.00	86.51	86.82	**87.68**	86.93	86.73	86.47	86.75
2 vs 7	97.97	99.42	99.32	99.25	99.34	99.37	99.42	**99.49**
3 vs 4	97.21	96.93	**97.41**	96.33	97.11	97.21	94.40	97.18
3 vs 5	94.02	**95.25**	94.59	92.12	94.42	94.51	90.52	94.79
3 vs 6	94.03	**95.18**	94.52	93.85	94.36	94.52	90.15	94.41
3 vs 7	98.80	91.62	**99.14**	96.10	97.13	96.81	98.48	96.49
4 vs 5	96.09	**96.72**	96.56	94.72	96.20	95.65	94.94	95.79
4 vs 6	95.41	95.15	**97.11**	96.95	96.56	96.45	96.72	96.76
4 vs 7	98.60	98.40	98.92	95.40	**98.99**	98.35	97.97	98.88
5 vs 6	**94.67**	94.24	94.39	94.38	94.12	94.52	91.99	94.36
5 vs 7	99.49	99.54	99.71	**99.88**	99.66	99.77	99.50	99.84
6 vs 7	**99.75**	99.66	97.42	96.49	99.32	99.56	99.61	99.27
Average	96.39	96.17	**96.60**	95.92	96.47	96.41	95.26	96.46
Std. Dev.	2.86	3.10	2.89	2.81	2.88	2.87	3.61	2.92

Table 2: Percentages of *Accuracy* estimated using a 4-fold cross-validation procedure.

Dataset	LR	RL	Back	Fwd	RFE	Relief	biPLS	*All*
1 vs 2	752.65	759.81	1729.15	**741.04**	770.18	1227.11	67912.28	*22.57*
1 vs 3	270.30	**270.21**	1811.58	496.27	366.82	495.38	39436.95	*10.73*
1 vs 4	**281.43**	323.26	920.03	1124.01	418.00	577.88	43381.10	*11.80*
1 vs 5	285.05	237.76	700.37	560.24	**219.91**	339.85	40846.62	*12.41*
1 vs 6	**274.96**	280.79	505.70	578.01	423.58	617.76	46180.82	*11.97*
1 vs 7	**37.37**	43.90	2074.44	884.15	315.38	458.88	40571.56	*9.77*
2 vs 3	**317.67**	398.95	1444.33	839.40	465.47	547.45	41187.64	*11.55*
2 vs 4	**273.57**	397.29	502.19	615.86	392.42	521.64	44000.70	*13.27*
2 vs 5	**461.43**	465.87	1363.58	766.46	468.65	578.78	39899.70	*16.15*
2 vs 6	547.67	558.44	613.07	990.51	**477.52**	620.56	46737.50	*17.17*
2 vs 7	**314.29**	335.55	1319.40	410.63	377.78	511.38	42215.36	*10.06*
3 vs 4	104.36	**81.01**	456.31	146.06	130.23	94.57	14441.13	*3.91*
3 vs 5	106.84	**95.77**	524.48	442.30	142.92	105.45	15102.25	*4.23*
3 vs 6	80.69	**79.66**	475.69	145.75	147.48	114.58	17108.40	*4.83*
3 vs 7	28.44	**21.81**	678.55	282.76	113.89	68.33	12438.09	*3.21*
4 vs 5	113.57	**82.90**	245.55	244.06	148.51	100.07	17686.72	*4.53*
4 vs 6	68.94	**53.74**	647.78	211.31	165.03	137.97	20051.43	*5.35*
4 vs 7	**31.92**	55.11	717.20	439.78	135.28	82.00	15430.84	*4.26*
5 vs 6	131.31	107.46	539.97	516.27	133.21	**100.48**	18713.01	*5.66*
5 vs 7	**17.86**	20.72	878.70	454.85	125.51	94.07	14804.12	*3.32*
6 vs 7	31.12	47.09	900.21	**17.88**	144.14	121.91	14210.64	*4.09*
Average	**215.78**	224.62	907.06	519.41	289.61	357.91	31064.61	*9.09*
Std. Dev.	189.41	200.07	499.64	287.32	172.28	289.79	15477.36	5.27

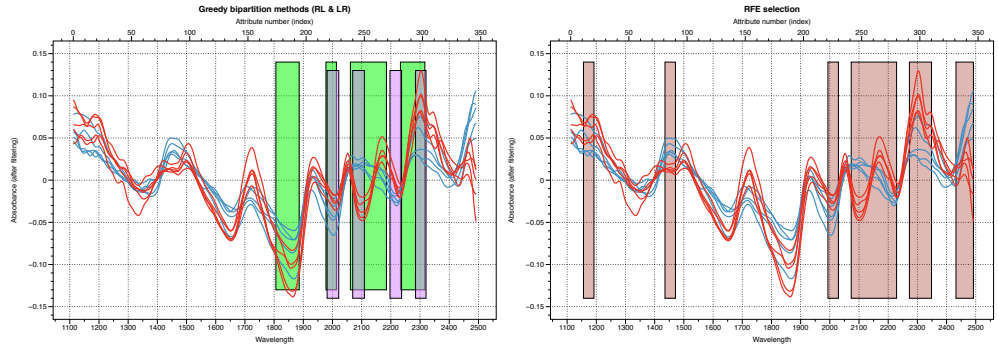
Table 3: Running time in seconds needed by each algorithm to obtain a single model from the full dataset.

	Discarded features / Intervals							*Quality*						
Dataset	LR	RL	Back	Fwd	RFE	Relief	biPLS	LR	RL	Back	Fwd	RFE	Relief	biPLS
1 vs 2	53/4	96/9	30/4	246/7	60/6	20/3	89/7	13.25	10.67	7.50	**35.14**	10.00	6.67	12.71
1 vs 3	99/3	140/7	90/5	236/7	176/8	40/5	89/6	33.00	20.00	18.00	**33.71**	22.00	8.00	14.83
1 vs 4	87/4	54/3	26/3	136/9	50/4	10/2	111/10	**21.75**	18.00	8.67	15.11	12.50	5.00	11.10
1 vs 5	97/1	130/3	20/3	226/8	20/3	30/3	87/9	**97.00**	43.33	6.67	28.25	6.67	10.00	9.67
1 vs 6	97/3	87/3	10/2	216/5	20/3	70/5	106/9	32.33	29.00	5.00	**43.20**	6.67	14.00	11.78
1 vs 7	282/3	249/2	286/6	16/2	0/1	220/7	97/6	94.00	**124.50**	47.67	8.00	0.00	31.43	16.17
2 vs 3	97/3	74/5	50/5	190/12	10/2	30/2	97/8	**32.33**	14.80	10.00	15.83	5.00	15.00	12.13
2 vs 4	151/7	43/4	10/2	216/9	50/6	30/3	101/6	21.57	10.75	5.00	**24.00**	8.33	10.00	16.83
2 vs 5	43/3	55/4	40/4	206/7	40/5	50/2	103/8	14.33	13.75	10.00	**29.43**	8.00	25.00	12.88
2 vs 6	22/2	11/2	10/2	176/9	0/1	0/1	98/5	11.00	5.50	5.00	19.56	0.00	0.00	**19.60**
2 vs 7	88/5	43/3	60/7	236/9	70/6	10/2	98/4	17.60	14.33	8.57	**26.22**	11.67	5.00	24.50
3 vs 4	86/5	130/6	70/7	226/9	50/5	10/2	113/8	17.20	21.67	10.00	**25.11**	10.00	5.00	14.13
3 vs 5	97/4	108/5	70/7	96/8	40/5	10/2	100/7	**24.25**	21.60	10.00	12.00	8.00	5.00	14.29
3 vs 6	118/3	108/2	60/6	236/10	20/3	10/2	92/8	39.33	**54.00**	10.00	23.60	6.67	5.00	11.50
3 vs 7	271/5	270/3	170/7	140/6	110/7	170/7	118/8	54.20	**90.00**	24.29	23.33	15.71	24.29	14.75
4 vs 5	77/4	141/5	20/3	196/6	20/3	80/4	94/7	19.25	28.20	6.67	**32.67**	6.67	20.00	13.43
4 vs 6	164/3	237/5	76/7	216/8	186/7	130/5	102/6	**54.67**	47.40	10.86	27.00	26.57	26.00	17.00
4 vs 7	228/3	226/5	130/7	76/5	86/6	90/5	94/7	**76.00**	45.20	18.57	15.20	14.33	18.00	13.43
5 vs 6	97/6	150/7	46/5	106/9	26/3	40/3	129/11	16.17	**21.43**	9.20	11.78	8.67	13.33	11.73
5 vs 7	303/4	261/4	290/6	36/2	240/6	230/6	111/7	**75.75**	65.25	48.33	18.00	40.00	38.33	15.86
6 vs 7	238/3	280/5	180/6	326/2	50/6	270/6	93/8	79.33	56.00	30.00	**163.00**	8.33	45.00	11.63
							Average	**40.21**	35.97	14.76	30.01	11.23	15.72	14.28
							Std. Dev.	27.62	28.90	12.46	30.95	8.82	11.84	3.26

Table 4: Quality scores of the intervals selected by each method in each binary learning task, as defined in (6). The number of features discarded/intervals selected are also indicated on the left side of the table.

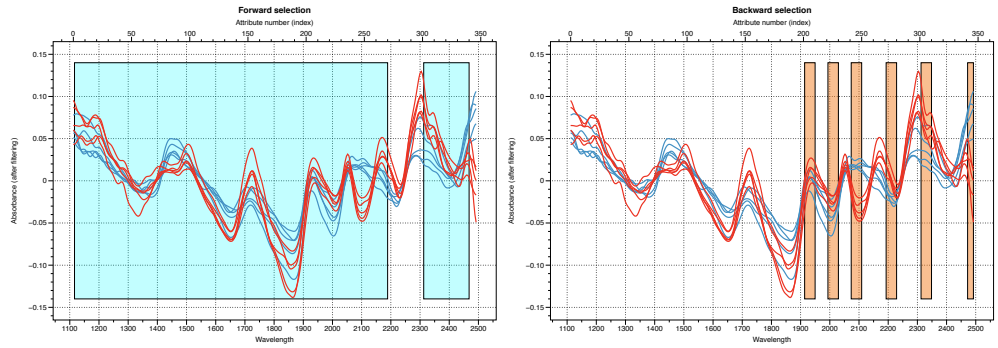
Method	Accuracy	#Disc. feat.	#Intervals	Quality	Running time
LR	4.48	3.07	2.81	**1.95**	**1.81**
RL	4.33	2.81	3.43	2.57	1.86
Back	**3.05**	5.10	3.98	5.38	5.67
Fwd	5.29	**2.29**	5.38	2.62	4.67
RFE	3.71	5.57	3.74	5.76	3.38
Relief	4.62	5.57	**2.69**	5.00	3.62
biPLS	6.90	3.60	5.98	4.71	7.00
All	3.62	–	–	–	–

Table 5: Average ranking positions for accuracy, number of discarded features, number of selected intervals, quality, and running time in seconds. The ranking position was computed from the results of the cross-validation, except for the running time, which was computed on a single training experiment with all data. The best result, i.e., the lowest in each column, is highlighted.



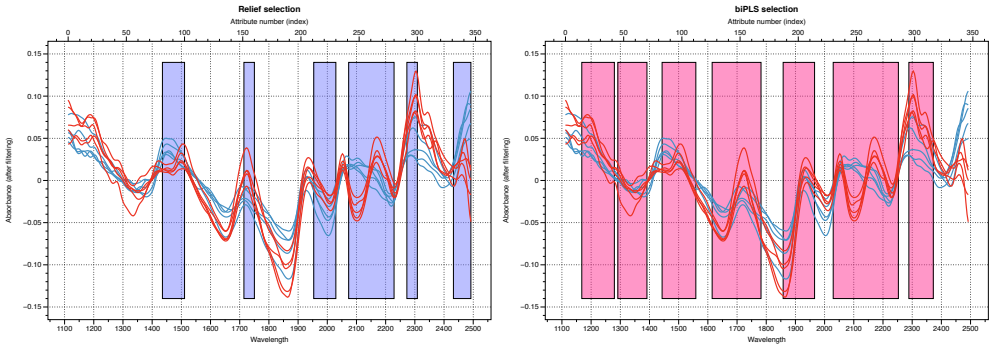
(a)

(b)



(c)

(d)



(e)

(f)

Figure 4: An example of intervals selected for discriminating by-products (class 5, blue spectra) and banned ingredients (class 7, red spectra). For the same training data, graph (a) depicts the intervals selected by *RL* (green) and by *LR* (purple); the rest of the graphs show the selection made by: (b) *RFE*, (c) *Fwd*, (d) *Back*, (e) *Relief* and (f) *biPLS*. These graphs also show 5 randomly-selected spectra of each class.