Improving the Discriminatory Power of a Near-Infrared Microscopy Spectral Library with a Support Vector Machine Classifier

V. FERNÁNDEZ-IBÁÑEZ, T. FEARN, E. MONTAÑÉS, J. R. QUEVEDO, A. SOLDADO, and B. DE LA ROZA-DELGADO*

Department of Animal Nutrition, Grasslands and Forages. Regional Institute for Research and Agro-Food Development. SERIDA, PO Box 13, 33300 Villaviciosa, Spain (V.F.-I., A.S., B.d.l.R.-D.); Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK (T.F.); Department of Computer Science, University of Oviedo, Campus de Viesques, Gijón 33204, Spain (E.M.); and Artificial Intelligence Center, University of Oviedo, Campus de Viesques, Gijón 33204, Spain (J.R.Q.)

A multi-group classifier based on the support vector machine (SVM) has been developed for use with a library of 48 456 spectra measured by near-infrared reflection microscopy (NIRM) on 227 samples representing 26 animal feed ingredients and 4 possible contaminants of animal origin. The performance of the classifier was assessed by a five-fold cross-validation, dividing at the sample level. Although the overall proportion of misclassifications was 27%, almost all of these involved the confusion of pairs of similar ingredients of vegetable origin. Such confusions are unimportant in the context of the intended use of the library, which is the detection of banned ingredients in animal feed. The error rate in discrimination between permitted and banned ingredients was just 0.17%. The performance of the SVM classifier was substantially better than that of the K-nearest-neighbors method employed in previous work with the same library, for which the comparable error rates are 36% overall and 0.39% for permitted versus banned ingredients.

Index Headings: **Near-infrared microscopy; NIR reflection microscopy; Spectral libraries; Support vector machine; Classification; Ingredients; Animal feeds.**

INTRODUCTION

Analytical methods that allow the identification and/or quantification of ingredients in animal feedstuffs are an essential part of an integrated food safety policy.[1] The detection of banned ingredients such as meat and bone meal (MBM) is of particular importance, both for the industry and for the regulatory authorities. The official analytical method to obtain information about the composition of feedstuffs is classical microscopy. In particular, optical microscopy is at the moment the only official method in the European Union for the detection of constituents of animal origin in feeds.[2]

There exist numerous works that research the use of the near-infrared (NIR) region for the analysis of ingredients in feed samples. The role of macro near-infrared spectroscopy for the detection and quantification of ingredients and possible contaminants in animal feed has been investigated by several authors.[3,4]

In recent years however, near-infrared reflection microscopy (NIRM) has been proposed as an alternative technology for the analysis of ingredients in feed samples, including the detection of banned processed animal proteins.[5–7] NIRM is an objective, sensitive, and highly selective technique. It combines the analytical advantages of microscopy and spectroscopy techniques.[6,8–11] NIRM is based on the collection of spectra from particles or very small areas ($\leq$50 μm) of a sample using a Fourier transform near-infrared reflection (FT-NIR) instrument attached to a microscope. The great advantages of this technique are that identification is not dependent on the expertise of the analyst and that it is possible to automate the procedure, increasing the number of samples analyzed per unit of time compared with classical microscopy while retaining the sensitivity advantage of microscopy.[11]

There are two major steps in the implementation of NIRM for an application such as this one. The first is the construction of a library of reference spectra, and the second is the development of an automated algorithm for comparing spectra from the sample of interest with those in the library in order to identify or possibly even quantify the ingredients in the sample.

Building the library is tedious and time consuming. Multiple spectra must be measured on multiple samples of each of many ingredients in order to represent the natural variability in the population of interest. Piraux and Dardenne[12] and Baeten et al.[13] describe the construction of spectral libraries including thousands of spectra of single particles from animal, vegetal, and mineral feed ingredients and the use of these to identify the origin of unknown particles. The work described here uses a library, also including thousands of spectra, in which the spectra were taken from very small areas of thin films of sample spread on a plate, rather than from individual particles. The optimization of measuring conditions and the construction of the library have been described in detail elsewhere.[11]

The main issue for the development of an automatic identification algorithm is that there is considerable natural variation in the spectra from any one ingredient, arising from the heterogeneity of the ingredients themselves. Given a spectrum from an unknown sample that happens to be soya meal, there will typically not be a single spectrum of soya meal in the library that provides a perfect match. It should, however, be true that the unknown spectrum is very close to at least some of the soya spectra in the library, and within the cluster defined by these spectra. The task for the identification algorithm is to find the similar spectra, or identify the appropriate cluster.

Classifying the spectrum of an unknown sample by comparing it with a library or database of known spectra is a problem that has many names, among them qualitative analysis, classification, discriminant analysis, product identification, and pattern recognition, and even more possible solutions.[14] In earlier work with the library used here, the K-nearest-neighbors (KNN) approach was employed.[10] This approach, which has the advantage of being very simple both to describe and implement, searches the library to find the K (5

Received 22 May 2009; accepted 26 October 2009.
* Author to whom correspondence should be sent. E-mail: broza@serida.org.

0003-7028/10/6401-0066$2.00/0

TABLE I. Ingredients, samples, and total number of NIRM spectra included in the spectral library.

Ingredient	ns	nsp	Ingredient	ns	nsp	Ingredient	ns	nsp
Lucerne dehydrated	24	5109	Sunflower seed	7	1578	Wheat middlings	1	210
Maize silage	23	4836	Oats	4	847	Citrus pulp	3	663
Fababean silage	2	450	Wheat	15	3210	Mineral correctors	6	1222
Grass silage	13	2713	Barley	20	4216	By-pass fat	3	630
Grass hay	2	421	Rye	5	1146	Milk powder	1	234
Cereal straw	27	5644	Maize	21	4826	Whey powder	1	210
Beet pulp	14	3039	Bran	2	419	Blood meal	1	235
Cotton seed	7	1016	Corn flakes	1	210	Meat and bone meal	4	937
DDGS Barley	1	202	Peas	2	427	Hemoglobin	1	234
Soybean meal	14	3120	Palm seed	1	218	Animal plasma	1	234

[a] ns: number of samples; nsp: number of spectra.

here) spectra most similar to the unknown and classifies the unknown according to which ingredient(s) these correspond to. In an assessment by cross-validation on the library, KNN gave excellent results for the identification of banned ingredients but was much less successful in the less important task of discriminating between ingredients such as barley and oats. The work reported here is an attempt to improve on the discriminatory performance of KNN by using an alternative approach taken from machine learning: a support vector machine (SVM).

The SVM method is one of the most popular machine learning approaches,[15] and versions of this method have been used successfully by other researchers working with near-infrared spectroscopy.[9,16,17]A major advantage in the context of a large library is that they are good at handling a large amount of data while consuming few resources, compared to, for instance, methods that generate a decision tree[18] or those based on artificial neural networks.

The objective of the present work was to develop a classification algorithm for NIRM spectra of feedstuff ingredients using the binary SVM classifier together with a voting strategy to cope with the multiplicity of groups and to assess its performance by cross-validation on the library.

EXPERIMENTAL

Samples. The 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 largest feed industries and rendering plants in the north of Spain in the framework of project RTA2005-00212-C02-00 from the INIA (National Institute of Agro-food Research) from 2005 to 2008, thus representing the variability encountered in the real production process. The samples were ground to a particle size of 1 mm prior to NIRM analysis.[11] The library comprised 48456 spectra measured on 227 samples representing 26 permitted and 4 banned ingredients. The ingredients included and the numbers of samples and spectra are listed in Table I and show a relative proportion akin to that encountered in the feed industry.

Near-Infrared Reflection Microscopy Analysis. An auto-image microscope connected to a Perkin-Elmer Spectrum One Fourier transform near-infrared (FT-NIR) spectrometer in reflection mode (1112–2500 nm) was used to measure the spectra. The sample was spread on a sample holder as a continuous film, an area in the center of the sample was selected and focused on, and spectra were measured using

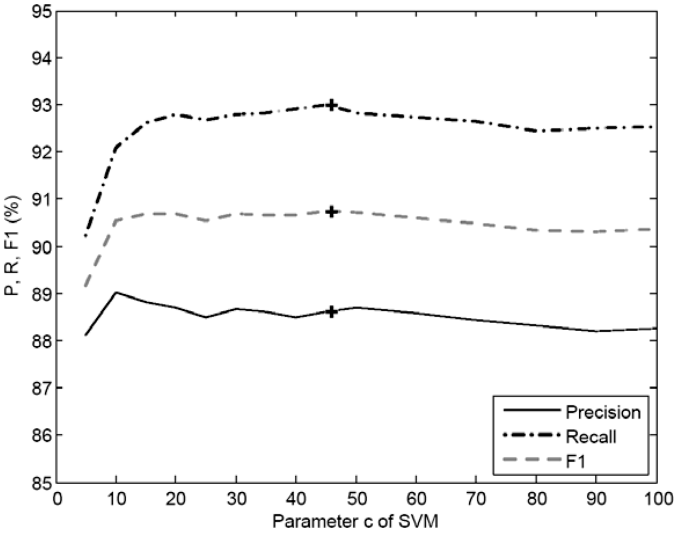

FIG. 1. Error rate of precision (P), recall (R), and combined measure F_1 as the parameter c of the SVM varies over the range from 0 to 100.

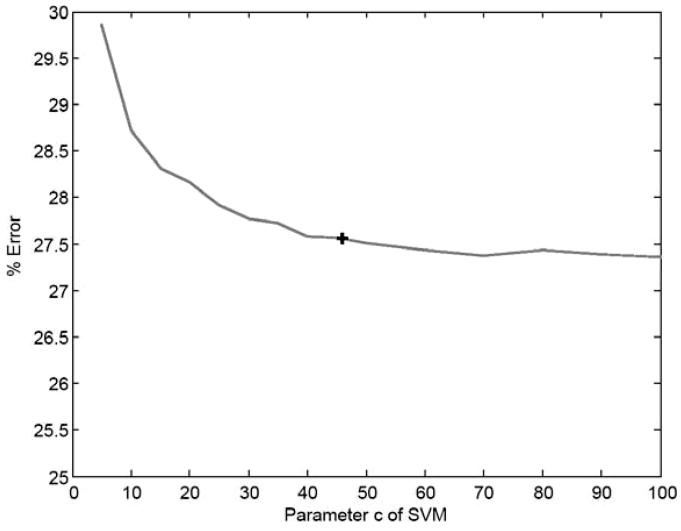

FIG. 2. Overall percentage classification error versus SVM's parameter c over the range from 0 to 100.

TABLE II. (a) Cross-validated confusion matrix for SVM applied to the library.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lucerne dehydrated	1	**3677**	170	15	396	10	634	54	29	21	19	14	2	1	14
Maize silage	2	351	**3575**	5	230	0	550	5	3	0	2	1	16	6	54
Fababean silage	3	103	11	**137**	42	6	76	2	6	0	8	17	2	1	2
Grass silage	4	454	410	3	**1546**	7	232	23	15	3	1	2	1	0	1
Grass hay	5	120	11	1	27	**139**	105	0	2	0	1	2	2	0	3
Cereal straw	6	555	105	11	117	9	**4614**	86	14	5	0	14	28	1	29
Beet pulp	7	162	10	2	6	0	211	**2587**	4	1	0	6	3	3	15
Cottonseed	8	94	37	0	38	0	64	20	**722**	1	17	0	7	1	6
DDGS Barley	9	19	10	0	13	0	1	0	0	**156**	0	2	0	0	0
Soybean meal	10	45	1	0	0	0	0	2	42	0	**2997**	3	1	3	4
Sunflower seed	11	73	1	0	2	0	38	2	0	1	13	**1437**	0	0	0
Oats	12	5	76	0	2	0	56	3	5	0	3	0	**529**	2	122
Wheat	13	3	2	0	1	0	0	2	0	0	2	0	2	**2191**	805
Barley	14	16	31	0	0	1	84	5	2	0	4	1	61	345	**2940**
Rye	15	2	6	0	0	0	0	1	0	0	1	0	1	59	817
Maize	16	4	9	0	0	0	3	2	0	0	3	0	20	267	846
Bran	17	32	15	0	0	0	3	6	6	0	9	1	11	9	45
Corn flakes	18	0	0	0	0	0	0	0	0	0	1	0	0	27	36
Peas	19	0	3	0	0	0	0	1	0	0	4	0	0	14	29
Palm seed	20	7	0	0	0	0	0	7	0	0	0	1	2	0	3
Wheat middlings	21	20	6	0	1	0	12	5	0	0	4	0	5	0	23
Citrus pulp	22	17	16	0	6	0	1	18	0	0	1	0	0	0	1
Mineral correctors	23	28	8	6	1	0	397	4	0	43	20	38	0	4	144
By-pass fat	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Milk powder	25	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Whey powder	26	1	0	0	1	0	0	0	0	0	12	0	0	0	0
Blood meal	27	0	0	0	0	0	0	0	0	0	0	2	0	0	0
Meat and bone meal	28	1	0	0	0	0	0	0	1	0	3	0	0	0	0
Hemoglobin	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Animal plasma	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE II. (b)

Summary	(1–26)	(27–30)
Permitted (1–26)	46743	73
Banned (27–30)	7	1633

fields of view of size 50 μm × 50 μm arranged in a 13 × 18 grid over this area, thus collecting approximately 200 spectra per sample. This method avoids any subjective selection of individual particles while still representing the inherent variability in the sample. As in previous works in NIRM,[6,8] spectra were obtained from the ratio between raw spectra and a Spectralon reference, and the spectral information was stored as $\log(1/R)$, recorded at 4 nm intervals over the range 1112–2500 nm after conversion from cm^{-1} using the Perkin Elmer software Spectrum v. 5.01.[19] Each recorded spectrum was the average of 70 scans to reduce the noise in the spectral data.[11]

Data Treatment. Before input to the data analysis, the spectral range was cut to 1288 to 2448 nm, in order to remove noisy regions at both ends of the range. The only other pretreatment applied for the results reported here was the subtraction of a linear trend line fitted by least squares to each individual spectrum. As an alternative, a first-derivative pretreatment was investigated, as was smoothing of the spectra.

Classification by Support Vector Machine. The SVM method is a universal binary classifier able to find linear or nonlinear threshold functions that optimize appropriate performance measures[15] in separating examples of one class (positive examples) from another class (negative examples). SVM builds the following function which classifies an unknown example with feature vector x, here the spectrum, as belonging to either the positive ($y = +1$) or negative ($y = -1$) class:

$$s = \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b$$
$$y = \begin{cases} +1 & \text{if } s > 0 \\ -1 & \text{else} \end{cases} \quad (1)$$

Here m is the number of training examples, $\{(x_i, y_i) | x_i \in R^n \wedge y_i \in \{+1, -1\}, i = 1, \ldots, m\}$ is the training set of examples, in this case the spectral library, $K(.,.)$ is a kernel function that measures the similarity of two feature vectors and can be used to introduce nonlinearity into the classifier, α_i are a set of coefficients that together define the linear function w, and b is a constant offset. Two kernels were investigated, the linear kernel, for which $K(.,.)$ is the simple dot product of the two vectors, and the radial basis function (RBF) kernel,[20] which is one of the most commonly used nonlinear alternatives.

The construction of the above classifier function involves maximizing the margin of error, which is based on the Structural Minimization Risk principle from computational learning theory.[21] This margin is defined as the distance from the function to the nearest positive and negative examples, which in fact form the support vectors. Maximizing this margin of error involves solving the following minimization problem.

$$minimize \ \frac{1}{2}\langle w, w\rangle + c\sum_{i=1}^{m}\xi_i$$
$$s.t. \ \forall i \, y_i[\langle w, h(x_i)\rangle + b] \geq 1 - \xi_i \quad \xi_i > 0 \quad (2)$$

TABLE II. (a) Extended.

	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Lucerne dehydrated	1	2	2	0	0	0	1	15	31	0	1	0	0	0	0	0
Maize silage	3	13	7	0	6	0	2	5	1	0	0	1	0	0	0	0
Fababean silage	4	2	1	0	6	1	3	2	8	0	2	6	2	0	0	0
Grass silage	1	0	0	0	1	0	0	8	4	0	0	1	0	0	0	0
Grass hay	1	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0
Cereal straw	1	3	1	0	1	0	0	0	45	1	0	1	1	2	0	0
Beet pulp	4	2	6	0	1	0	1	4	11	0	0	0	0	0	0	0
Cottonseed	1	2	1	1	0	1	0	1	0	0	0	0	0	2	0	0
DDGS Barley	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Soybean meal	0	0	3	0	3	0	0	4	0	0	0	0	0	12	0	0
Sunflower seed	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
Oats	0	34	10	0	0	0	0	0	0	0	0	0	0	0	0	0
Wheat	21	142	14	14	8	0	0	0	3	0	0	0	0	0	0	0
Barley	113	583	12	6	3	0	0	1	8	0	0	0	0	0	0	0
Rye	**43**	213	1	0	1	0	0	0	1	0	0	0	0	0	0	0
Maize	87	**3559**	14	8	3	0	0	1	0	0	0	0	0	0	0	0
Bran	0	3	**277**	0	1	0	0	0	1	0	0	0	0	0	0	0
Corn flakes	4	57	0	**85**	0	0	0	0	0	0	0	0	0	0	0	0
Peas	0	0	0	0	**375**	0	0	0	1	0	0	0	0	0	0	0
Palm seed	0	0	0	0	0	**198**	0	0	0	0	0	0	0	0	0	0
Wheat middlings	0	2	5	0	0	0	**124**	3	0	0	0	0	0	0	0	0
Citrus pulp	0	0	0	0	6	0	0	**380**	0	0	215	1	0	0	0	1
Mineral correctors	0	4	1	0	10	0	1	0	**452**	9	0	0	0	52	0	0
By-pass fat	0	0	0	0	0	0	0	0	0	**630**	0	0	0	0	0	0
Milk powder	0	0	0	0	0	0	0	160	0	0	**71**	0	0	0	0	1
Whey powder	0	0	0	0	0	0	0	1	0	0	0	**195**	0	0	0	0
Blood meal	0	0	0	0	0	0	0	0	0	0	0	0	**233**	0	0	0
Meat and bone meal	0	0	0	0	0	0	0	0	0	0	0	0	0	**928**	0	4
Hemoglobin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**234**	0
Animal plasma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**234**

where $\langle .,. \rangle$ is the dot product, $h(.)$ is a function implicitly defined by the kernel $K(.,.)$, each ξ_i is the distance allowed as a tolerance between the separating hyperplane defined by the linear function w and the ith example, and c is a crucial parameter that penalizes over-fitting and helps to ensure the generalizability of the classifier. A wide range of values for this parameter was examined in the tuning process.

Although the SVM has been shown to perform fast and well with many features,[15] their main disadvantage is that they just solve binary problems, i.e., discrimination between two classes. One way of overcoming this limitation when there are $n > 2$ ingredients is to convert the original problem into a set of n binary problems, each one consisting of deciding whether a spectrum comes from a certain ingredient or not. This process is known as one-against-the-rest approach.[22] Another alternative is the one-against-one[23] approach, which entails bringing face to face every pair of ingredients. This approach results in as many binary problems as the number of combinations of n elements taken in twos, that is, $(n^2 - n)/2$. In the former, the number of binary problems to solve is lower than in the latter, but it is easier for the SVM classifier to separate one ingredient from another than one ingredient from a mixture of the rest.[24] Since the library contains just $n = 30$ ingredients, the number of problems to solve is not unreasonably high if the one-against-one approach is adopted, and this was the approach used. Once a set of functions separating each pair of ingredients is built, it is necessary to combine them to obtain a global function able to predict the ingredient from a spectrum. A common approach to build a global model is called voting.[25] This involves applying each of the pair-wise functions in turn to the spectrum. Each time, the winner is assigned a point, and in the end the algorithm predicts the ingredient that has accumulated the most points.[26]

The algorithm used in this study was libsvm.[27] It is embedded in the Spider[28] Matlab toolbox.

Cross-Validation. A cross-validation with five folds and one repetition was carried out on the library in order to choose the parameter c and to assess the performance of the classifier.[29] It consists of splitting the training set in five folds or segments and obtaining five models, each one using four from the five folds as the training set and the remaining fold as a test set. The five performance measures thus calculated were then averaged. The cross-validation was performed as far as possible at the sample level. This means that in splitting the spectra into the five folds, all the spectra from one sample were put in the same fold. In the cases for which this was not possible, because the ingredient had fewer than five samples, the splitting was carried out in a way that separated the samples as much as possible.

Performance Measures and Tuning. Several measures of performance were considered when tuning c. As well as the overall error rate, three measures of performance that focus on the discrimination between permitted and banned ingredients were considered.[30] The precision, P, is the proportion of all the spectra classified as coming from a banned (i.e., animal) ingredient that actually do come from such an ingredient. The recall, R, is the proportion of all the spectra truly coming from a banned animal ingredient that are correctly classified as such. The third measure, F_1, is a balanced combination of P and R that is commonly adopted as an appropriate trade-off between the two. These measures are most commonly used when there is one target class having priority over others, as is the case

TABLE III. (a) Cross-validated confusion matrix for KNN applied to the library.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lucerne dehydrated	1	**3058**	538	44	188	34	846	167	61	24	20	41	5	4	12
Maize silage	2	196	**4132**	5	73	8	229	13	9	2	4	3	15	10	79
Fababean silage	3	236	26	0	38	6	95	12	13	0	7	4	1	3	0
Grass silage	4	633	1084	24	**469**	35	360	23	46	2	9	6	3	2	1
Grass hay	5	193	40	5	35	**2**	113	2	1	0	6	9	3	0	3
Cereal straw	6	765	297	14	126	34	**4099**	163	23	1	10	5	25	4	22
Beet pulp	7	142	38	0	26	3	275	**2444**	4	3	1	4	19	13	38
Cottonseed	8	109	39	0	18	5	70	35	**657**	0	26	0	23	1	4
DDGS barley	9	22	9	0	0	0	2	1	1	**144**	6	16	1	0	0
Soybean meal	10	43	3	1	0	4	4	5	85	1	**2690**	69	4	4	7
Sunflower seed	11	106	15	5	3	7	21	10	3	6	168	**1228**	0	0	1
Oats	12	5	170	0	2	1	32	50	11	1	0	1	**150**	7	242
Wheat	13	2	8	0	2	0	2	3	2	0	2	0	11	**1783**	1027
Barley	14	3	98	0	3	2	69	18	1	1	8	1	111	251	**2698**
Rye	15	1	7	0	0	0	3	2	0	0	1	0	13	58	775
Maize	16	2	21	0	0	0	5	7	0	0	5	0	26	477	865
Bran	17	20	21	0	2	0	5	16	7	4	8	0	65	9	100
Corn flakes	18	0	0	0	0	0	0	0	0	0	0	0	1	25	61
Peas	19	0	0	0	0	1	0	0	0	0	1	0	0	23	125
Palm seed	20	6	0	0	0	0	2	9	0	1	1	1	1	0	1
Wheat middlings	21	4	9	0	0	0	14	3	0	0	0	0	23	2	32
Citrus pulp	22	27	25	1	10	0	4	101	1	0	0	0	5	1	8
Mineral correctors	23	33	3	0	1	0	15	2	0	1	6	10	6	9	122
By-pass fat	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Milk powder	25	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Whey powder	26	1	1	0	0	0	0	0	0	0	8	3	0	0	0
Blood meal	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Meat and bone meal	28	0	0	0	0	0	0	0	2	0	10	0	0	0	0
Hemoglobin	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Animal plasma	30	0	0	0	0	0	0	0	0	0	2	0	0	0	0

TABLE III. (b)

Summary	(1–26)	(27–30)
Permitted (1–26)	46644	172
Banned (27–30)	15	1625

here with the detection of animal ingredients. The following expressions define them in terms of true and false positives and negatives.

$$P = \frac{TP}{TP + FP} \tag{3a}$$

$$R = \frac{TP}{TP + FN} \tag{3b}$$

$$\frac{1}{F_1} = \frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R} \tag{3c}$$

where TP is the number of spectra correctly classified as coming from a banned animal ingredient, FP is the number of spectra incorrectly classified as coming from a banned animal ingredient, and FN is the number of spectra classified as coming from a permitted ingredient, when in fact they come from a banned animal ingredient.

RESULTS AND DISCUSSION

Comparing the various data pretreatments investigated, the alternatives of subtracting a linear baseline or transforming to first derivative gave similar results, and the baseline subtraction was chosen for its simplicity. Smoothing the spectra improved their appearance, but not the classification performance. Thus all the results reported below use a linear baseline subtraction only as data pretreatment.

Comparing the two kernels used, the best results were obtained with the linear kernel, and it is these results that are described below. The success of this simple kernel suggests that this particular problem is in fact linearly separable.

Figures 1 and 2 show the overall error rate, the precision P, recall R, and combined measure F_1 when the c parameter of the SVM ranges from 0 to 100. Although the error rates in Fig. 2 are high, it will be seen later, when the errors are analyzed in

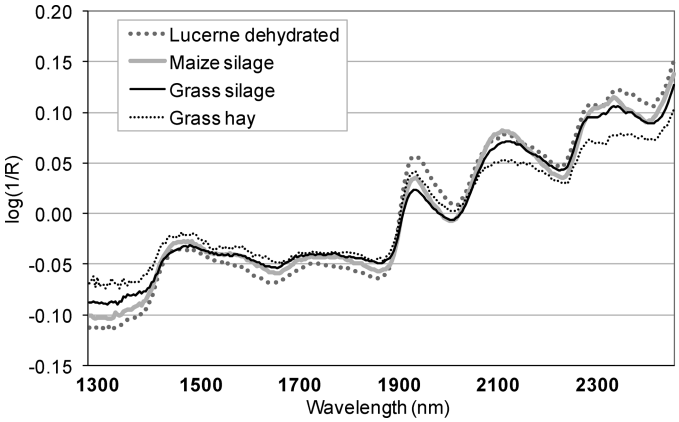

FIG. 3. Mean raw spectra of four forages: Lucerne dehydrated, maize silage, grass silage, and grass hay.

TABLE III. (a) Extended.

	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Lucerne dehydrated	0	6	6	0	0	0	0	39	13	0	0	3	0	0	0	0
Maize silage	13	35	1	1	1	1	0	5	1	0	0	0	0	0	0	0
Fababean silage	0	7	0	0	0	0	0	0	1	0	0	1	0	0	0	0
Grass silage	0	2	0	0	0	0	0	6	7	0	0	1	0	0	0	0
Grass hay	0	0	0	0	0	0	0	0	5	0	0	4	0	0	0	0
Cereal straw	0	0	3	0	1	0	1	6	40	0	0	0	0	0	0	5
Beet pulp	2	8	6	0	0	0	2	9	1	0	0	1	0	0	0	0
Cottonseed	0	4	2	0	0	0	0	11	0	0	0	0	0	12	0	0
DDGS barley	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Soybean meal	1	13	3	0	8	0	1	0	3	0	0	16	0	155	0	0
Sunflower seed	0	1	0	0	0	0	0	0	0	0	0	4	0	0	0	0
Oats	19	129	8	0	0	0	9	2	5	0	0	3	0	0	0	0
Wheat	108	225	7	3	20	0	0	0	5	0	0	0	0	0	0	0
Barley	404	463	19	0	34	0	2	0	27	0	0	3	0	0	0	0
Rye	**169**	111	0	0	3	0	0	0	2	0	0	1	0	0	0	0
Maize	222	**3165**	7	1	14	0	0	0	8	1	0	0	0	0	0	0
Bran	2	69	**53**	2	0	5	29	1	1	0	0	0	0	0	0	0
Corn flakes	7	116	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Peas	2	15	0	0	**229**	0	0	0	24	0	0	7	0	0	0	0
Palm seed	0	2	0	0	0	**194**	0	0	0	0	0	0	0	0	0	0
Wheat middlings	3	16	18	1	0	0	**77**	0	6	0	0	2	0	0	0	0
Citrus pulp	6	1	0	0	1	0	0	**239**	2	0	231	0	0	0	0	0
Mineral correctors	4	9	4	0	31	0	4	0	**942**	16	0	4	0	0	0	0
By-pass fat	0	0	0	0	0	0	0	0	0	**630**	0	0	0	0	0	0
Milk powder	0	1	0	0	0	0	0	231	0	0	0	0	0	0	0	0
Whey powder	0	0	0	0	1	0	0	1	0	0	0	**195**	0	0	0	0
Blood meal	0	0	0	0	0	0	0	0	0	0	0	0	**235**	0	0	0
Meat and bone meal	0	0	0	0	0	0	0	0	0	1	0	0	0	**924**	0	0
Hemoglobin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**234**	0
Animal plasma	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**232**

detail, that many of these errors are confusions between similar cereals, for instance between wheat and barley. For simplicity though, these overall measures were used for tuning. The SVM's c parameter controls the trade off between generalization and over-fitting. In order to avoid both extreme situations, an exhaustive search for an adequate value of such a parameter was performed using cross-validation, taking F_1 as the target measure. In light of these figures, the parameter c was fixed at 45, since this value maximizes R and F_1 while the overall error rate decreases only slightly for higher values of c.

The cross-validated confusion matrix for the parameter choice $c = 45$ is given in Table II (a). So that the results may be compared, the confusion matrix obtained in the earlier work[10] using KNN is given in Table III (a). The results are broadly similar, in that in both cases there are many confusions between pairs such as maize silage and grass silage, or barley and rye, but relatively few between ingredients of cereal origin and ingredients of animal origin. The confusions between similar ingredients are neither surprising nor particularly important. For example, it is clear from the mean spectra of the four forages in Fig. 3 that these ingredients have spectrally very similar shapes. They are also functionally similar in the context of animal feed. The critical discrimination is that between the first 26 ingredients in the tables, which are all permitted in animal feed, and the last four, which are banned. Thus, the important areas of the tables are the last four rows and columns, where the number of confusions is low for both approaches.

The summary tables in Tables II (b) and III (b) are the result of collapsing the full tables to the two categories "permitted" and "not permitted". It can be seen that the SVM approach improves significantly on the earlier KNN results, with the number of important errors being approximately halved. In both cases the most common false negative (animal classified as vegetable) is meat and bone meal wrongly identified as soya. Interestingly, the most common false positive (vegetable classified as animal) is different for the two approaches. KNN classifies 155 soya spectra as meat and bone meal. SVM only makes 12 errors of this type, but classifies 52 spectra of mineral correctors as meat and bone meal. These confusions are related to the mineral composition of bones present in meat and bone meals.

CONCLUSION

The aim of the work reported here was to improve on the performance of the KNN classifier previously developed for use with the spectral library of feed ingredients. The SVM approach was successful in reducing the overall proportion of classification errors from the 36% of the KNN approach to 27% on vegetal origin ingredients, but more importantly it substantially reduced the number of confusions between permitted and banned ingredients. Of the 48 456 classifications made, the number of errors of banned ingredients was only 80, an error rate of 0.17%, compared with 187 (0.39%) for the KNN approach. In the context of the proposed use of NIRM for the detection of banned ingredients, this is a significant improvement.

With this study it was proved that the combination of NIRM with SVM as a chemometric classification technique should allow a regulatory laboratory to certify the presence of meat and bone meal in animal feed with an error rate lower than 0.2%.

ACKNOWLEDGMENTS

This work was supported by the Spanish project RTA2005-00212-C02-00 from the INIA; MEC and FEDER grant TIN2007-61273. The authors are grateful to INIA for a pre-doctoral fellowship of Mª del Valle Fernández Ibáñez and post-doctoral financial support of Dr. Ana Soldado. Thanks are also given to Alfonso Carballal and Nutrition Laboratory staff for their technical assistance.

1. European Union. Regulation (EC) No.178/2002. Off. J. Eur. Comm. L **31,** 1–24 (2002).
2. European Union. Regulation (EC) No.152/2009. Off. J. Eur. Comm. L **54,** 1–130 (2009).
3. D. Pérez-Marín, A. Garrido-Varo, J. E. Guerrero-Ginel, and A. Gómez-Cabrera, Animal Feed Sci. Technol. **116,** 333 (2004).
4. D. Pérez-Marín, A. Garrido-Varo, and J. E. Guerrero, Appl. Spectrosc. **60,** 1432 (2006).
5. L. W. D. van Raamsdonk, C. von Holst, V. Baeten, G. Berben, A. Boix, and J. de Jong, Animal Feed Sci. Technol. **133,** 73 (2007).
6. B. de la Roza-Delgado, A. Soldado, A. Martinez-Fernández, F. Vicente, A. Garrido-Varo, D. Pérez-Marín, M. J. de la Haba, and J. E. Guerrero-Ginel, Food Chem. **105,** 1164 (2007).
7. M. J. De la Haba, J. A. Fernández-Pierna, O. Fumière, A. Garrido-Varo, J. E. Guerrero, D. C. Pérez-Marín, P. Dardenne, and V. Baeten, J. Near Infrared Spectrosc. **15,** 81 (2007).
8. V. Baeten, C. von Holst, A. Garrido, J. Vancutsem, A. M. Renier, and P. Dardenne, Anal. Bioanal. Chem. **382,** 149 (2005).
9. C. von Holst, V. Baeten, A. Boix, B. Slowikowski, J. A. Fernández-Pierna, S. Tirendi, and P. Dardenne, Anal. Bioanal. Chem. **392,** 313 (2008).
10. V. Fernández-Ibáñez, T. Fearn, A. Soldado, and 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 Chem., paper submitted (2009).
11. M. V. Fernández-Ibáñez, A. Soldado, F. Vicente, A. Martínez-Fernández, and B. de la Roza-Delgado, J. Near Infrared Spectrosc. **16,** 43 (2008).
12. F. Piraux and P. Dardenne, "Feed authentication by near-infrared microscopy", in *Near Infrared Spectroscopy: Proceedings of the 9th International Conference,* A. M. C. Davies and R. Giangiacomo, Eds. (NIR Publications, Chichester, UK, 2000), p. 535.
13. V. Baeten, A. Michotte Renier, G. Sinnaeve, and P. Dardenne, "Analyses of feedingstuffs by near-infrared microscopy (NIRM): detection and quantification of meat and bone meal (MBM)", in *Proceedings of the 6th International Symposium on Food Authenticity and Safety* (Nantes, 2001), p. 1.
14. T. Naes, T. Isaksson, T. Fearn, and T. Davies, *Multivariate Calibration and Classification* (NIR Publications, Chichester, UK, 2002), p. 221.
15. T. Joachims, *Making Large-Scale Support Vector Machine Learning Practical Advances in Kernel Methods: Support Vector Machines*, B. Scholkopf, C. Burges, and A. Smola, Eds. (MIT Press, Cambridge, MA, 1998).
16. D. Perez-Marin, A. Garrido-Varo, J. E. Guerrero, T. Fearn, and A. M. C. Davies, Appl. Spectrosc. **62,** 536 (2008).
17. R. P. Cogdill and P. Dardenne, J. Near Infrared Spectrosc. **12,** 93 (2004).
18. J. R. Quinlan, *C4.5. Programs for Machine Learning,* M. Kaufmann, Ed. (Morgan Kauffman Publishers, San Francisco, 1993).
19. *IR Spectroscopy Software: User's Guide* (Perkin-Elmer Instruments LLC, UK, 2002).
20. S. Lukaszyk, Comput. Mech. **33,** 299 (2004).
21. V. N. Vapnik, *The Nature of Statistical Approaches to Text Categorization* (Springer-Verlag, New York, 1995).
22. B. Schölkopf, C. Burges, and V. Vapnik, "Extracting Support Data for a Given Task", in *1st International Conference on Knowledge Discovery and Data Mining* (Menlo Park, Canada, 1995).
23. O. Chapelle, P. Haffner, and V. N. Vapnik, *Support Vector Machines for Histogram-Based Image Classification IEEE Transactions on Neural Networks*, **10**(5), (1999).
24. J. Fürnkranz, J. Machine Learning Res. **2,** 721 (2002).
25. V. Koltchinskii, D. Panchenko, and F. Lozano, "Further explanation of the effectiveness of voting methods: The game between margins and weights", in *Proceedings 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory* (Amsterdam, The Netherlands, 2004), p. 241.
26. J. Fürnkranz, "Round Robin Rule Learning", in *Proceedings of the 18th International Conference on Machine Learning (ICML-01)* (Warsaw, Poland, 2001), p. 146.
27. C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines (2001), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
28. J. Weston, A. Elisseeff, G. Bakir, and F. Sinz, http://www.kyb.tuebingen.mpg.de/bs/people/spider/.
29. T. Dietterich, Neural Comput. **10,** 1895 (1998).
30. F. Sebastiani, ACM Computing Survey **34**(1), (2002).