Collaborative representation based classifier with partial least squares regression for the classification of spectral data

Weiran Song[a], Hui Wang[a], Paul Maguire[b], Omar Nibouche[a]

[a]School of Computing, [b]School of Engineering, Ulster University, BT37 0QB, Newtownabbey, Co. Antrim, UK

ABSTRACT

The need to classify high-dimensional spectral data is an increasingly common occurrence in rapid and non-destructive detection of object features and chemical species using spectroscopy. Partial least squares discriminant analysis (PLS-DA) is an effective method for spectral data classification, which is based on a multivariate regression model. Although powerful, PLS-DA suffers from performance degradation under complex conditions such as nonlinearity, imbalance and multiclass, which are common in real applications. Collaborative representation-based classifier (CRC) is a new machine learning algorithm which represents a query by a linear combination of training samples and classifies the query based on the representation. It offers the possibility of classifying even under nonlinearity, imbalance and multiclass conditions. In this paper, we present a novel method for spectral data classification, namely CRC-WPLS, which reaps the benefits of both PLS regression and CRC. This method searches for a weighted, linear combination of all training samples to represent the query by using PLS regression, and then assigns the query to the class which yields the least approximation error. CRC-WPLS is compared to PLS-DA, kernel PLS-DA, support vector machine (SVM), random forest (RF) and representation-based classifiers on fourteen general machine learning datasets and three spectral datasets. Experimental results show the proposed method can outperform 7 baseline methods in most cases, and achieve a high classification accuracy (> 90%) for low grade spectra obtained from portable instrumentation.

Keywords: Classification, Partial least squares, Collaborative representation, Spectral data.

1. Introduction

The combination of spectroscopy and chemometrics provides an effective tool for identifying the chemical compositions of a material in many fields such as food, pharmaceutical and biomedical science. It aims to reveal the qualitative or quantitative relationship between the high-dimensional spectra and corresponding identities by means of a classification model. Recently, the utilization of low-cost and portable spectrometers is gaining increasing attention for many field-based applications. However, variable

environmental conditions and inevitable instrument limitations pose serious challenges for implementation of field-based or portable strategies at the required level of accuracy, precision and cost. Resultant spectra suffer with considerable noise and variability and classifying this data using conventional chemometric methods such as partial least squares discriminant analysis (PLS-DA) may lead to significant performance degradation.

PLS-DA is an adaptation of PLS regression for the classification problem. It searches for independent *latent variables* (LVs), that can be used to effectively predict the response. Typically, when high dimensionality and high collinearity are present in small-sample data, the regression coefficient of PLS is stable compared to that of ordinary least squares (OLS) and can be computed efficiently. Thus, PLS is practically suitable for spectral data analysis. For absorption spectra, according to the Beer-Lambert law, there is a linear relationship between the absorbance and object properties such as analyte concentrations and optical depth [1]. However, under non-ideal conditions such as stray light, detector-based and chemical-based effects, nonlinear variations may be introduced into absorption spectral data [2, 3]. Under such conditions, linear PLS models generally degrade in performances [4, 5]. Other spectra data, such as non-equilibrium plasma emission in the visible region, are inherently non-linear.

Many attempts have been made to improve the prediction performance with nonlinear spectral data. Kernel PLS (KPLS) is a popular approach [6]. It maps data into Hilbert feature space, where a linear PLS model is built. The nonlinear relationship among variables in the original data space becomes linear in the feature space after such mapping [6] (the Covers theorem), so KPLS can effectively describe nonlinear data and hence has potential to improve the prediction performance. One disadvantage of KPLS is that it is difficult to attribute the performance of the model to specific variables in the original data space [7]. Also, if the dataset has inadequate samples, kernel methods are prone to overfitting [5].

Machine learning algorithms can be used to analyse spectral data. Actually many machine learning algorithms have similar or even better performance than PLS-DA on various spectral data classification tasks, for example, random forest (RF) [8] on material identification, artificial neural network (ANN) and support vector machine (SVM) on food discrimination [9, 10] . Nevertheless, these algorithms remain less preferable than PLS-DA because they are not good at identifying significant variables nor at revealing the interaction between variables [11]. While developing high accuracy models is of critical importance in chemometrics, it is also necessary to provide information on the scope and applicability of such models to real world measurement conditions. Instrumental and environmental sensitivities are often non-linear and ill-characterised, so that even repeat measurements under nominally similar conditions can produce widely different spectra, resulting in nonlinear data. This reduces the confidence that any given training set can be used reliably under naturally varying measurement conditions. Generally, the variable sensitivity to instrument and environment factors is known or can be determined only for a small subset of variables. Hence by clearly mapping the model (i.e. variable) sensitivities to determinable physical sensitivities, appropriate

training protocols can be matched to application conditions. Many traditional machine learning algorithms do not consider this important implementation factor, while PLS provides a good insight into the causes of discrimination via weights and loadings [12].

Previously we investigated the performance of local PLS-DA with non-Euclidean distance on low and high dimensionality data [13] and we have demonstrated the value of a nearest clusters based approach (NCPLS-DA) for explicitly addressing multimodality and nonlinearity issues in spectral data classification [14]. Recently, representation-based classifiers have been widely investigated in non-spectral data, such as face recognition [15, 16] and hyperspectral image identification [17, 18]. Among these classifiers, collaborative representation classifier (CRC) [16] is a standard one that uses a linear combination of all training samples to represent a query and attributes the query to the class which yields the least approximation error. Ideally, the approximation of the class to which a query belongs most closely resembles the query. By comparing all approximations, it is possible to directly see the class-wise disparity for each variable, which cumulatively controls the classifier decision. This allows a simple and clear interpretation of how variables contribute to a classification model for a query. To our knowledge, representation-based classifiers have not been investigated for the classification of spectral data. It is therefore of value to consider the application of representation-based classification for spectral data and determine its predictive accuracy.

In this paper, we present a new method for the classification of spectral data, which combines PLS regression and CRC, namely CRC-PLS. This method constructs a global approximation for a given query based on all training samples by using PLS regression and then divides the global approximation into several independent approximations according to classes. The query will be assigned to the class which provides the most accurate approximation. To improve the prediction performance, we apply a weighting scheme in PLS regression which is based on the distance between the query and all training samples. Therefore, training samples which are in the neighbourhood of the query will provide a higher contribution to the global approximation. This method, termed CRC-WPLS, has been tested on public machine learning and spectral datasets which covers highly complex data structures such as high dimensionality, multiclass and imbalance. In order to test the capability of CRC-WPLS more fully, we also created our own dataset from near infrared (NIR) reflectance spectra of apples obtained from a low-cost portable NIR spectrometer under uncontrolled field conditions.

The remainder of the paper is organized as follows: Section 2 briefly reviews the related works and describes the proposed method. Datasets description and experimental settings are given in Section 3. Section 4 presents and discusses the classification results on seventeen datasets. Conclusions are drawn in Section 5.

2. Theory and algorithm

In this paper, scalars are defined as lower case characters, vectors are in bold lowercase characters and matrices as bold uppercase characters. Superscripts t and −1 represent transpose and inverse operations, respectively. Let **X** be the $n \times p$ data matrix (rows corresponding to samples and columns to variables) which contains c-classes, and $\mathbf{X}_i$ denote the data matrix of the ith class. The response vector and matrix are denoted as **y** and **Y**, respectively.

2.1. Partial least squares discriminant analysis (PLS-DA)

PLS is a standard method for processing a wide spectrum of chemical data problems, which relies on the basic assumption that the investigated system or process is driven by a set of underlying LVs (also called latent vectors, score vectors, or components). It extracts LVs by projecting both **X** and **Y** onto a subspace such that the pairwise covariance between the LVs of **X** and **Y** is maximized. To ensures the mutual orthogonality of the LVs, this procedure is iteratively carried out by using deflation scheme [19] which subtracts from **X** and **Y** the information explained by their rank-one approximations based on score vectors. Nonlinear iterative partial least squares (NIPALS) [20] and SIMPLS [21] are two widely used PLS algorithms. When the number of LVs is not high, SIMPLS provides similar performance with less computational cost compared to NIPALS [22]. This paper adopts SIMPLS algorithm, which can be summarized as follows:

(1) Calculate **s** as

$$\mathbf{s} = \mathbf{X}^{\mathrm{T}} \mathbf{y}$$

(2) Calculate the quantities **r** (PLS weights for **X**), **t** (PLS scores for **X**), **p** (PLS loadings for **X**) and q (PLS loading for **y**) as follows:

$$\mathbf{r} = \mathbf{s}$$

$$\mathbf{t} = \mathbf{X}\,\mathbf{r}$$

$$\mathbf{t} = \mathbf{t} / \| \mathbf{t} \|$$

$$\mathbf{r} = \mathbf{r} / \| \mathbf{r} \|$$

$$\mathbf{p} = \mathbf{X}^{\mathrm{T}} \mathbf{t}$$

$$q = \mathbf{y}^{\mathrm{T}} \mathbf{t}$$

(3) Store **r**, **t**, **p** and q in **R**, **T**, **P** and **q**, respectively.

(4) Update **s** as

$$\mathbf{s} = \mathbf{s} - \mathbf{P}\,(\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}\,\mathbf{P}^{\mathrm{T}}\mathbf{s}$$

(5) Go to step (2) to calculate the next latent variable until reaching required number of latent variables

(6) Calculate the regression vector as

$$\mathbf{b} = \mathbf{R}\,\mathbf{q}$$

PLS-DA is an adaption of PLS2 algorithm for the classification purpose, which transforms the categorical vector into numerical responses using dummy matrix coding [23]. A sample is then assigned to the class corresponding to the maximum value in the dummy vector.

2.2. Representation-based classification

Representation-based classification relies on the underlying assumption that a $1 \times p$ query vector $\mathbf{x}_q$ can be linearly represented by all training samples or within-class training samples. Such representation is based on regression which can be varied in optimizations. CRC and nearest regularized subspace (NRS) [17] are two typical methods which respectively uses l_2-norm and distance-weighted Tikhonov regularization [24] to handle the small-sample problem in representing the query. Suppose the columns of $\mathbf{X}^{\mathrm{T}}$ is normalized to have unit l_2-norm, CRC uses all training samples concurrently to code the query via l_2-norm regularized least squares, which can be expressed as

$$\arg\min_{\boldsymbol{\alpha}} \left\|\mathbf{x}_q^{\mathrm{T}} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}\right\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2\,, \tag{1}$$

where λ is a regularization parameter, and $\boldsymbol{\alpha}$ is an $n \times 1$coefficient vector. Taking derivative with regard to $\boldsymbol{\alpha}$ and setting the resultant equation to zero yields

$$\boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{x}_q^{\mathrm{T}}\,, \tag{2}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix. Then CRC partitions the collaborative representation $\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}$ according to the classes and calculates the residual of the ith approximation as

$$e_i = \left\|\mathbf{x}_q^{\mathrm{T}} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_i\right\|_2\,, \tag{3}$$

where $\boldsymbol{\alpha}_i$ is the partitioned coefficients corresponding to the ith class in $\boldsymbol{\alpha}$. The class label of the query is decided as

$$\mathrm{class}(\mathbf{x}_q) = \arg\min_{i=1,\cdots,c}(e_i)\,. \tag{4}$$

Besides the approximation residual e_i, the l_2-norm $\| \boldsymbol{\alpha}_i \|_2$ also contains some discrimination information between classes [16]. To improve the classification accuracy, CRC with regularized least square (CRC-RLS) classifies the query as

$$\text{class}(\mathbf{x}_q) = \arg \min_{i=1,\cdots,c} (e_i / \|\boldsymbol{\alpha}_i\|_2) \,. \tag{5}$$

NRS generates an approximation of the query independently from all available training samples per class and uses a distance-weighted Tikhonov regularization to enhance the distinction between classes. The objective function of NRS is

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{x}_q^{\mathrm{T}} - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\alpha}_i \right\|_2^2 + \lambda \|\boldsymbol{\Gamma}_i \boldsymbol{\alpha}_i\|_2^2 \,, \tag{6}$$

where the Tikhonov matrix $\boldsymbol{\Gamma}_i$ is a diagonal matrix specific to the ith class for the query in the form of

$$\boldsymbol{\Gamma}_i = \begin{bmatrix} \left\| \mathbf{x}_q^{\mathrm{T}} - \mathbf{x}_{i,1}^{\mathrm{T}} \right\|_2 & & 0 \\ & \ddots & \\ 0 & & \left\| \mathbf{x}_q^{\mathrm{T}} - \mathbf{x}_{i,n_i}^{\mathrm{T}} \right\|_2 \end{bmatrix}, \tag{7}$$

where $\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,n_i}$ are the rows of $\mathbf{X}_i$. From (6) and (7), the samples which are dissimilar to the query will give less contribution toward the approximation than samples which are close to the query. Likewise, (6) also has a closed-form solution, which is

$$\boldsymbol{\alpha}_i = (\mathbf{X}_i \mathbf{X}_i^T + \lambda \boldsymbol{\Gamma}_i^T \boldsymbol{\Gamma}_i)^{-1} \mathbf{X}_i \mathbf{x}_q^{\mathrm{T}} \,. \tag{8}$$

Then the class assignment of the query is calculated according to (4).

2.3. Collaborative representation-based classification with partial least squares (CRC-PLS)

The CRC-PLS method is proposed which combines collaborative representation with PLS regression for the data classification. This method firstly searches for a linear combination of all training samples to code a query via PLS regression. Then the obtained regression coefficients are partitioned according to the class labels and used to form the approximation of each class. Finally, the query is attributed to the class which provides the least approximation error.

In addition to CRC and sparse representation-based classification (SRC) [15], CRC-PLS can effectively code the query if the small-sample problem exists in representation phase. However, it has been reported that the l_1-norm sparsity constraint in SRC does not truly improves the classification performance [16]. In fact, traditional CRC and SRC ignore the distance relationship between the query and all training samples when coding the query, resulting in unsatisfied classification accuracies [17, 25]. To distinguish the contribution

of samples toward the query, we improve the CRC-PLS via a variable-weighted approach [26], which is expressed as

$$\mathbf{X}_{\mathbf{w}}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathrm{diag}(\mathbf{w}) , \tag{9}$$

where $\mathbf{w}$ is an $n \times 1$ weighting vector with all the elements being non-negative values and diag($\mathbf{w}$) is an $n \times n$ matrix whose diagonals are the elements of $\mathbf{w}$. When $\mathbf{w}$ is assigned with discrete values 0 and 1, this process becomes variable selection for $\mathbf{X}^{\mathrm{T}}$ (sample selection for $\mathbf{X}$) which can remove uninformative training samples with respect to the query. In this paper, we adopt continuous values in the range (0, 1] to represent the significance of samples as follows [27]:

$$w_i = \exp\left(-\frac{\varphi d_i}{\sigma_d}\right) \tag{10}$$

$$d_i = \sqrt{\left(\mathbf{x}_{\boldsymbol{i}} - \mathbf{x}_{\boldsymbol{q}}\right)^{\mathrm{T}}\left(\mathbf{x}_{\boldsymbol{i}} - \mathbf{x}_{\boldsymbol{q}}\right)} , \tag{11}$$

where w_i is the ith element of $\mathbf{w}$, φ is a localization parameter, and σ_d is the standard deviation of $\{d_i\}$. Then the representation of the query can be obtained by calculating the regression coefficients of response $\mathbf{x}_q^{\mathrm{T}}$ on predictor variables $\mathbf{X}_{\mathbf{w}}^{\mathrm{T}}$. The proposed method is summarized in **Algorithm 1**.

Algorithm 1. CRC-WPLS.

Input: A data matrix of training samples $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{c} \in R^{n \times d}$ for c classes; a test query $\mathbf{x}_q \in R^d$; number of latent variable k; localization parameter φ.

1: Calculate the weighting vector $\mathbf{w}$ according to Eq. (10) and (11).

2: Generate the weighted matrix $\mathbf{X}_{\mathbf{w}}^{\mathrm{T}}$ according to Eq. (9).

3: Calculate the regression coefficients $\boldsymbol{\alpha}$ of $\mathbf{x}_q^{\mathrm{T}}$ on $\mathbf{X}_{\mathbf{w}}^{\mathrm{T}}$ using SIMPLS algorithm with k latent variables.

4: Partition $\boldsymbol{\alpha}$ into $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, …, $\boldsymbol{\alpha}_c$ according to the class labels and calculate the approximation error of the ith class

$$e_i = \left\|\mathbf{x}_q^{\mathrm{T}} - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_i\right\|_2 .$$

5: Predict the class label of $\mathbf{x}_q$

$$\hat{\mathbf{y}}_q = \arg\min_{i=1,\cdots,c}\left(e_i / \|\boldsymbol{\alpha}_i\|_2\right) .$$

Output: A class label $\hat{\mathbf{y}}_q$.

3. Experimental

3.1. Datasets

The performance of CRC-WPLS was tested on a collection of public machine learning and spectral datasets as well as one spectral dataset (NIR–apple) which we obtained experimentally to represent the challenge of low grade spectra obtained from portable instrumentation in uncontrolled conditions [28]. Fourteen machine learning datasets, including one spectral dataset (ARCENE), are mostly obtained from the UCI data repository [29] which cover a diversity of data structures including high dimensionality (ARCENE and Lung), multiclass (Leaf and Movement) and imbalance (ECOLI and Glass). Also, some datasets have been analysed in state-of-the-art sciences. The ARCENE dataset is generated by mass spectroscopy for cancer detection and contains 10,000 variables. Many of these variables are irrelevant and feature selection prior to classification is suggested [30]. We therefore apply the ReliefF algorithm [31] only on this dataset to remove 90% of the variables. The information about the 14 machine learning datasets are shown in Table 1.

Table 1

Information on 14 machine learning datasets

Datasets	Samples	Variables	Categories	Areas
ARCENE	200	1000	2	Mass spectrometry
Breast tissue	106	9	4	Impedance measurements
ECOLI	336	7	8	Protein localization sites
Forest types	523	27	4	Remote sensing
Glass	214	9	6	Physical
Ionosphere	351	34	2	Signal processing
Leaf	340	14	36	Image classification
Lung	73	325	7	Gene expression
Movement	360	90	15	Movement recognition
Parkinson's	195	22	2	Voice measurements
Seed	210	7	3	X-ray imaging techniques
Sonar	208	60	2	Sonar signals
SPECTF	267	44	2	SPECT heart images
Wine	178	12	3	Chemical analysis

Description of the three spectral datasets are as follows:

- NIR-apple [28]: a total of 182 apples were scanned in reflectance mode using a portable near infrared (NIR) Ocean Optics spectrometer. Each spectrum contains 512 variables in the wavelength range 901.06-1721.24 nm with an interval of 1.65 nm. There were two species of apples (Gala and Pink Lady) in this dataset, and each species had non-organic and organic samples. Since the apple species can be accurately identified, the task was to differentiate non-organic (96 samples) and organic (86 samples) apples.

- FTIR-oil [32]: this dataset was obtained by Fourier transform infrared (FTIR) spectroscopy in absorption mode with attenuated total reflectance sampling under controlled laboratory conditions. A total of 120 authenticated extra virgin olive oils (including duplicates) were used to distinguish the country of their origins: Greece, Italy, Portugal and Spain (respectively 20, 34, 16 and 50 samples of each). The wavelength of the spectra ranged from 798.89 to 1896.81 nm with an interval of 1.93 nm.
- FTIR-fruit [33]: a total of 983 mid infrared spectra were collected in two classes: 'strawberry' and 'non-strawberry' purees, respectively 351 and 632 of each class. Each spectrum contained 235 variables in 899.33-1802.56 nm with an interval of 3.86 nm taken under controlled laboratory conditions in absorption mode.

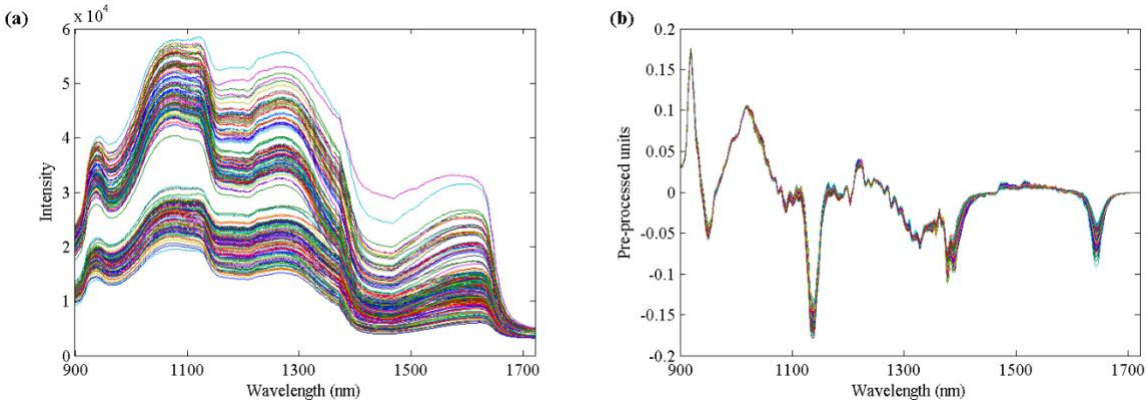

Fig. 1. Raw spectra (a) and pre-processed spectra (b) of NIR-apple dataset.

FTIR-oil and FTIR-fruit are two publicly available spectral datasets, which have been studied in many works [34, 35, 36]. To improve the performance of classification models, we directly apply the same pre-processing steps as in [34]: the raw data matrix was centred by subtracting the mean spectrum, scaled by standard deviation, and processed by the Savitzky-Golay first-order derivative (5-point moving window and second-order polynomial). The same pre-processing steps are also applied on NIR-apple dataset, the raw and pre-processed apple spectra are shown in Fig. 1.

3.2. Experimental settings

The proposed CRC-WPLS is compared to PLS-DA methods (PLS-DA and KPLS-DA), representation-based classifiers (CRC, CRC-RLS and NRS), SVM and RF. We use DUPLEX algorithm [37] to split each dataset into training and testing sets according to the ratio of 2:1, and then set proper ranges to tune the parameters of these classifiers on training data via 10-fold cross validation (machine learning and FTIR-fruit datasets) or leave-one-out cross validation (NIR-apple and FTIR-oil datasets).

The range of LVs in PLS is varied between 1 and 10 if the minimum number between n (sample) and p (variable) is above 10, otherwise, from 1 to min (n, p). The regularization parameter λ of CRC, CRC-RLS

and NRS is mostly set from 10^{-7} to 10^{2} on a logarithmic scale. The parameters of KPLS-DA (LV × σ), SVM (C × γ) and CRC-WPLS (LV × φ) are selected by grid search. The width of the radial basis function (RBF) σ in KPLS-DA is varied from 10^{-3} to 10^{5} on a logarithmic scale. The combination of the regularization parameter C (C = 1, 10, 100, 1000) and the RBF kernel parameter γ (γ = 0.01, 0.1, 1, 10) allows the construction of 16 SVM models in total for validation [38]. As the optimal value of the localization parameter φ is usually found in the range of 0 to 10 [39], we adjust φ to the values of 0.01, 0.05, 0.1, 0.5, 1, 5 and 10. The parameters of RF, i.e., the number of trees grown (*ntree*), the number of predictors sampled for splitting at each node (*mtry*) and *nodesize*, have been set to their default values (*ntree* = 500, *mtry* = $\sqrt{p}$ and *nodesize* = 1) [40]. It has been reported that such default setting often yields a good prediction model [41]. The optimal parameters of different algorithms on machine learning datasets are provided in Table 2.

Table 2

The optimal parameters of different algorithms for 14 machine learning datasets.

	PLS-DA	KPLS-DA		CRC	CRC-RLS	NRS
	LVs	LVs	log(σ)	log(λ)	log(λ)	log(λ)
ARCENE	6	9	3	-1	-1	0
Breast tissue	4	9	4	-5	-4	0
ECOLI	5	9	2	-1	-1	2
Forest type	10	8	2	-3	-2	1
Glass	8	10	0	-7	-3	0
Ionosphere	4	3	0	-3	0	0
Leaf	10	10	2	-6	-6	-1
Lung	5	10	2	0	0	0
Movement	10	10	1	-4	-3	-1
Parkinson's	8	7	4	-4	-6	0
Seed	5	9	4	-6	-5	0
Sonar	8	9	0	-1	-1	1
SPECTF	2	10	2	0	-2	2
Wine	6	7	4	-6	-6	-1

	SVM		CRC-PLS	CRC-WPLS	
	C	γ	LVs	LVs	φ
ARCENE	1000	0.01	8	2	1
Breast tissue	1000	0.01	8	5	0.5
ECOLI	10	1	5	6	0.1
Forest type	100	0.01	9	9	0.5
Glass	10	1	6	7	0.5
Ionosphere	10	0.01	9	5	0.01
Leaf	1000	1	10	10	1
Lung	1000	0.01	2	2	0.5
Movement	1000	0.01	10	8	1
Parkinson's	100	10	8	10	1
Seed	100	1	6	4	0.1
Sonar	10	0.1	9	3	0.5
SPECTF	10	1	4	1	0.5
Wine	1000	0.01	6	6	1

We also provide the classification performance of NC-PLSDA in comparison to the proposed method on spectral datasets. NC-PLSDA constructs a local PLS model using nearest clusters of a query, which has three parameters, LVs, clustering numbers (CN) and nearest clusters (NC). The CN can be directly obtained based on the average Euclidean distance between the mean of samples and the means of clusters, while the NC is empirically set and validated in a proper range [14]. The optimal parameters of different algorithms on spectral datasets are shown in Table 3.

Table 3

The optimal parameters of different algorithms for 3 spectral datasets.

	PLS-DA	KPLS-DA		NCPLS			CRC
	LVs	LVs	log(σ)	LVs	CN	NC	log(λ)
NIR-apple	9	10	0	6	34	8	-7
FTIR-oil	5	6	-3	2	32	8	-9
FTIR-fruit	10	10	-2	7	26	3	-9

	CRC-RLS	NRS	SVM		CRC-PLS	CRC-WPLS	
	log(λ)	log(λ)	C	γ	LVs	LVs	φ
NIR-apple	-4	-2	100	0.01	9	7	5
FTIR-oil	-9	1	10	0.01	6	4	0.5
FTIR-fruit	-9	0	10	0.1	10	8	5

4. Results and discussion

4.1. Results on UCI datasets

The classification results of nine algorithms for machine learning datasets are shown in Table 4. Among the nine algorithms, the proposed CRC-WPLS yields the best classification results in half of 14 datasets. From the results averaged overall datasets (last column of Table 2), the accuracy of CRC-WPLS respectively exceeds that of PLS-DA and CRC by over 10% and 7%, that potentially reveals the existence of nonlinearity. Moreover, the CRC-WPLS outperforms two well-preforming nonlinear algorithms, i.e., RF and SVM, by over 1.5% of classification accuracy, providing the most accurate results among the nine algorithms. It also presents high capacity in handling high-dimensional, multiclass and imbalance problems. Other representation-based classifiers can also outperform PLS-DA in most cases. Among these classifiers, the CRC, CRC-RLS and CRC-PLS achieve comparable results on average, however, the accuracies of these methods are below that of NRS. Furthermore, the optimal parameters λ of CRC and CRC-RLS tend to be the same, while the optimal LVs of CRC-WPLS is often less than or equal to that of PLS-DA (see Table 2).

Table 4

Classification accuracy (%) of different algorithms for machine learning datasets.

	ARCENE	Breast tissue	ECOLI	Forest type	Glass	Ionosphere	Leaf	Lung
PLS-DA	80.6	82.9	85.7	83.3	64.8	79.5	52.2	70.8

KPLS-DA	83.6	80	84.8	84.5	70.4	90.6	54.9	70.8
CRC	82.1	91.4	86.6	83.9	64.8	88	54	75
CRC-RLS	82.1	82.9	**87.5**	86.2	60.6	**94**	59.3	70.8
NRS	83.6	80	83.9	89.1	76.1	88.9	73.5	75
SVM	**88.1**	88.6	84.8	87.4	67.6	**94**	77	75
RF	83.6	85.7	**87.5**	86.2	**78.9**	92.3	**84.1**	75
CRC-PLS	82.1	82.9	84.8	85.6	67.6	93.2	54.9	**83.3**
CRC-WPLS	**88.1**	**94.3**	86.6	**92**	74.6	**94**	79.6	**83.3**

	Movement	Parkinson's	Seed	Sonar	SPECTF	Wine	*Average*
PLS-DA	48.3	87.7	94.3	76.8	85.4	94.9	77.7
KPLS-DA	47.5	83.1	95.7	81.2	86.5	93.2	79.1
CRC	60.8	84.6	95.7	78.3	**88.8**	89.8	80.3
CRC-RLS	76.7	83.1	95.7	81.2	**88.8**	96.6	81.8
NRS	**87.5**	81.5	92.9	76.8	86.5	94.9	83.6
SVM	83.3	**93.8**	92.9	81.2	84.3	96.6	85.3
RF	79.2	92.3	87.1	**87**	87.6	**98.3**	86.1
CRC-PLS	69.2	83.1	95.7	78.3	**88.8**	93.2	81.6
CRC-WPLS	86.7	87.7	**97.1**	82.6	**88.8**	94.9	**87.9**

4.2. Results on spectral datasets

We graphically present the selection of the optimal λ and LVs via leave-one-out cross validation on NIR-apple dataset, as in Fig. 2 a and b. The parameters σ and φ have been set to their corresponding optimal values (see Table 2), respectively in KPLS-DA and CRC-WPLS. The accuracy of three classical representation-based classifiers maintains stability at the first few regularization parameters and decreases until reaching the specific value of λ. Three of the PLS-based methods, PLS-DA, KPLS-DA and CRC-PLS have poor performance when the number of LVs is small, while CRC-WPLS always goes beyond these methods in each LV. We also demonstrate a grid search of the optimal LVs and φ for CRC-WPLS, which is shown in Fig. 2c, as a grayscale colormap. The CRC-WPLS obtains the highest result of 93.4% when LVs and φ equals to 7 and 5, respectively.

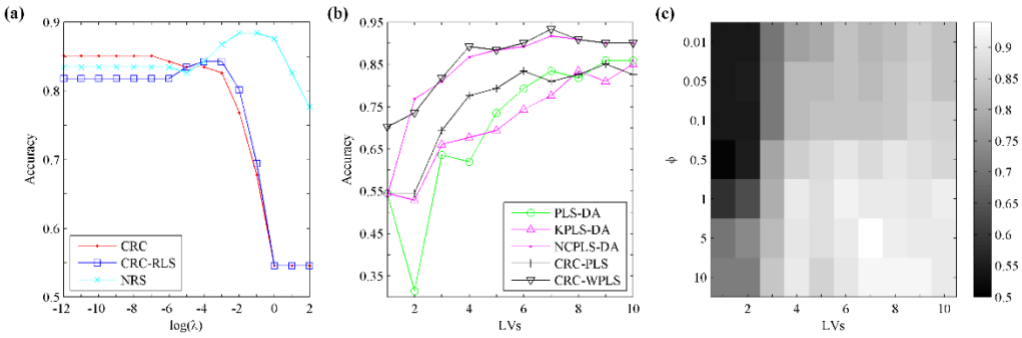

Fig. 2. Performance of eight algorithms on NIR-apple dataset evaluated by leave-one-out cross validation (a and b) and the overall validation results of CRC-WPLS with varying parameters (c).

The overall training and testing performance of different algorithms on spectral datasets are shown in Table 5. For NIR-apple dataset, CRC-WPLS achieves the highest results of 93.4% and 90.2%, respectively in validation and classification phases, which ranks the first among the ten algorithms. Other representation-based classifiers except NRS yield less accurate results compared to PLS-DA. NCPLS-DA outperforms PLS-DA by over 3% in testing phases, while SVM and RF provide the same results to PLS-DA. The CRC-WPLS and NRS respectively obtains the maximum accuracy of 96.7% and 87.1% in identifying non-organic and organic classes.

For FTIR-oil dataset, algorithms based on PLS-DA present good performance in validation while SVM and RF obtain the top accuracy (95%) in classification. The CRC-WPLS can also reach the highest validation accuracy of 96.3% but fails in classification by using the corresponding optimal parameters. However, if we select the parameters (LVs = 1 and φ = 5) corresponding to the third highest validation result (93.8%), the classification accuracy of CRC-WPLS will become to 92.5% which is identical to that of NCPLS-DA and CRC.

For a fair comparison between all algorithms, the same indices are used for 10-fold cross validation on FTIR-fruit dataset. The CRC-WPLS and NRS achieve the highest validation accuracy of 97.3% while CRC-RLS obtains the best classification result of 98.2%. The SVM and RF give comparable classification results to CRC-WPLS and NRS, which exceed PLS-DA by over 1.5%. Testing samples from the majority class (non-strawberry) can be successfully identified by NRS while those from the minority class (strawberry) are optimally recognised by CRC and CRC-RLS.

Table 5

Validation and classification accuracy (%) of different algorithms for spectral datasets.

NIR-apple	Training	Testing	Non-organic	Organic
PLS-DA	86	85.2	86.7	83.9
KPLS-DA	85.1	78.7	83.3	74.2
NCPLS-DA	91.7	88.5	93.3	83.9
CRC	85.1	68.9	76.7	61.3
CRC-RLS	84.3	77	83.3	71
NRS	88.4	**90.2**	93.3	87.1
SVM	90.1	85.2	93.3	77.4
RF	82.6	85.2	93.3	77.4
CRC-PLS	85.1	78.7	86.7	71
CRC-WPLS	**93.4**	**90.2**	96.7	83.9

FTIR-oil	Training	Testing	Greece	Italy	Portugal	Spain
PLS-DA	**96.3**	90	100	100	85.7	72.7
KPLS-DA	**96.3**	90	100	100	85.7	72.7
NCPLS-DA	93.8	92.5	100	100	85.7	81.8
CRC	95	92.5	100	100	85.7	81.8
CRC-RLS	92.5	90	100	100	71.4	81.8
NRS	92.5	87.5	85.7	100	85.7	72.7

SVM	**96.3**	**95**	100	100	100	81.8
RF	93.8	**95**	100	100	100	81.8
CRC-PLS	93.8	85	100	86.7	71.4	81.8
CRC-WPLS	**96.3**	85	85.7	93.3	85.7	72.7

FTIR-fruit	Training	Testing	Non-strawberry	Strawberry
PLS-DA	94.2	95.4	95.5	95.2
KPLS-DA	92.7	93	92.7	94
NCPLS-DA	96.6	96.6	97.1	95.2
CRC	95.6	96.6	95.9	98.8
CRC-RLS	96.5	**98.2**	98	98.8
NRS	**97.3**	97.3	100	89.2
SVM	96.9	97.6	98	96.4
RF	95.7	97	98.4	92.8
CRC-PLS	94.8	96	98.4	89.2
CRC-WPLS	**97.3**	97.3	98	95.2

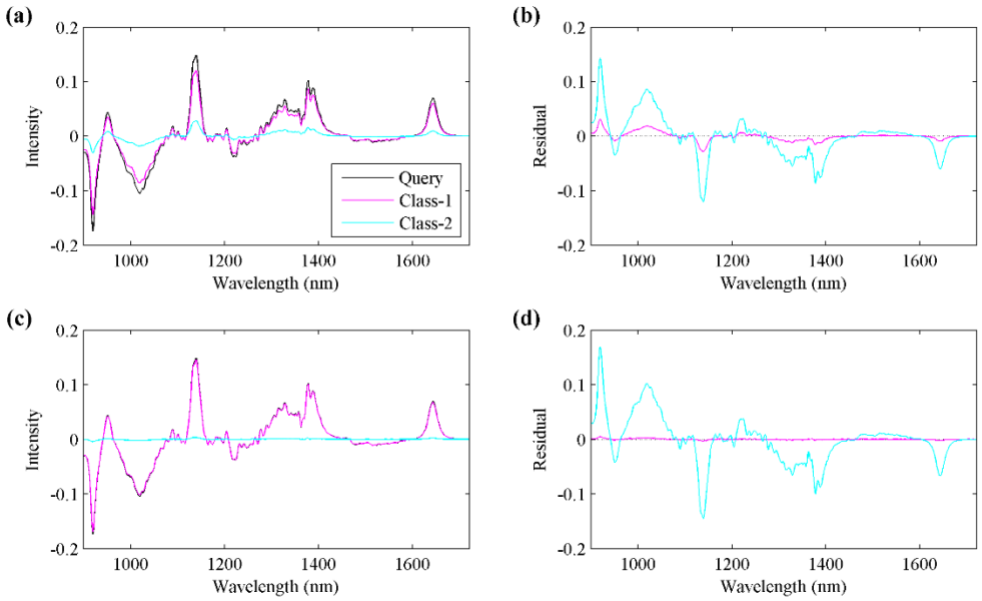

Fig. 3. The CRC (a) and CRC-WPLS (b) approximations of a test query in NIR-apple dataset. The corresponding residuals of CRC and CRC-WPLS in approximating are respectively given in (c) and (d).

We interpret the underlying mechanism of representation-based classifiers via an example on NIR-apple dataset. Samples are normalized to have unit l_2-norm. From the theory of CRC, the collaborative representation is partitioned into class-wise approximations to predict the label of the query. A query spectrum and its CRC approximations are shown in Fig. 3a. After subtracting the query from each approximation, the residual spectrum of non-organic class (class-1) approaches to zero values compare to that of organic class (class-2), as shown in Fig. 3b. Therefore, this query will be attributed to non-organic class by CRC. We also present the approximations and residuals by using CRC-WPLS as in Fig. 3c and d. CRC-WPLS provides a closer non-organic approximation to the query than CRC, which demonstrates the

improved efficiency of CRC-WPLS in distinguishing class-wise approximations. Furthermore, the residuals obtained by NRS are shown in Fig. 4. By comparing class-wise residuals, training samples from the non-organic class provides the most accurate approximation.

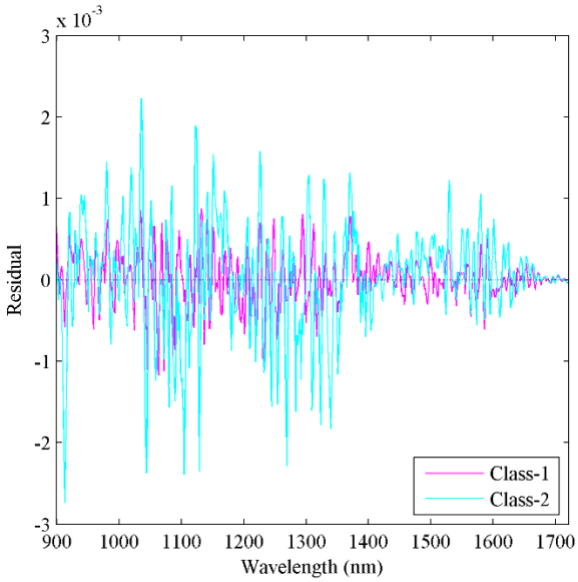

Fig. 4. The residuals of NRS in approximating a test query in NIR-apple dataset.

From the experimental results, the proposed CRC-WPLS outperforms PLS-DA in twelve out of 14 machine learning datasets and two out of 3 spectral datasets. Moreover, it yields better accuracies than two well-performing methods SVM and RF in most cases. A main reason for the outperformance is CRC-WPLS implements a weighting scheme according to the distance between the query and each training sample. When coding a representation, such weighting scheme enlarges the contribution of samples which are adjacent to the query, meanwhile, lessens the influence of sample which are dissimilar to the query. CRC-PLS adopts PLS regression to code the representation, which does not significantly improve the classification accuracies compared to CRC and CRC-RLS. NRS provides more accurate results than PLS-DA over half of the datasets, however, it will degrade in performance when datasets such as Lung and FTIR-oil have a limited number of training samples.

5. Conclusions

In this paper, we have proposed a new method termed CRC-WPLS for spectral data classification to improve the prediction performance of PLS-DA and representation-based classifiers. This method utilizes PLS regression with a weighting scheme to find the collaborative representation for a query, and then attributes the query to the class which yields the least approximation error. Through our experiments on benchmark datasets, we found that the proposed CRC-WPLS can result in higher classification performance than 7 baseline methods.

Representation-based classifiers provide a simple and intuitive interpretation on how each wavelength contribute to the classification decision of a query spectrum. Therefore, significant variables can be identified by comparing class-wise approximations. Our future work will develop variable selection approaches for spectral data analysis in terms of representation-based classification.

Reference

[1] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (2009) 1201–1222.

[2] Miller, Charles E. "Sources of non-linearity in near infrared methods." NIR news 4.6 (1993) 3-5.

[3] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing?, TrAC - Trends Anal. Chem. 50 (2013) 96–106.

[4] J. Peng, L. Guo, Y. Hu, K.F. Rao, Q. Xie, Maximum correntropy criterion based regression for multivariate calibration, Chemom. Intell. Lab. Syst. 161 (2017) 27–33.

[5] Despagne, Frédéric, D. Luc Massart, and Paul Chabot. "Development of a robust calibration model for nonlinear in-line process data." Analytical chemistry 72.7 (2000) 1657-1665.

[6] Rosipal, Roman, and Leonard J. Trejo. "Kernel partial least squares regression in reproducing kernel Hilbert space." Journal of machine learning research 2. Dec (2001) 97-123.

[7] Postma, G. J., P. W. T. Krooshof, and L. M. C. Buydens. "Opening the kernel of kernel partial least squares and support vector machines." Analytica chimica acta 705.1-2 (2011) 123-134.

[8] Zhang, Tianlong, et al. "Classification of steel samples by laser-induced breakdown spectroscopy and random forest." Chemometrics and Intelligent Laboratory Systems 157 (2016) 196-201.

[9] Cajka, Tomas, et al. "Recognition of beer brand based on multivariate analysis of volatile fingerprint." Journal of Chromatography A 1217.25 (2010) 4195-4203.

[10] Luna, Aderval S., et al. "Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 100 (2013) 115-119.

[11] Brereton, Richard G. Chemometrics for pattern recognition. John Wiley & Sons, 2009.

[12] Brereton, Richard G., and Gavin R. Lloyd. "Partial least squares discriminant analysis: taking the magic away." Journal of Chemometrics 28.4 (2014) 213-225.

[13] Song, Weiran, et al. "Local Partial Least Square classifier in high dimensionality classification." Neurocomputing 234 (2017) 126-136.

[14] Song, Weiran, et al. "Nearest clusters based partial least squares discriminant analysis for the classification of spectral data." Analytica chimica acta 1009 (2018) 27-38.

[15] Wright, John, et al. "Robust face recognition via sparse representation." IEEE transactions on pattern analysis and machine intelligence 31.2 (2009) 210-227.

[16] Zhang, Lei, Meng Yang, and Xiangchu Feng. "Sparse representation or collaborative representation: Which helps face recognition?" Computer vision (ICCV), 2011 IEEE international conference on. IEEE, 2011.

[17] Li, Wei, et al. "Nearest regularized subspace for hyperspectral classification." IEEE Transactions on Geoscience and Remote Sensing 52.1 (2014) 477-489.

[18] Cui, Minshan, and Saurabh Prasad. "Class-dependent sparse representation classifier for robust hyperspectral image classification." IEEE Transactions on Geoscience and Remote Sensing 53.5 (2015) 2683-2695.

[19] Höskuldsson, Agnar. "PLS regression methods." Journal of chemometrics 2.3 (1988) 211-228.

[20] Wold, Herman. "Nonlinear iterative partial least squares (NIPALS) modelling: some current developments." Multivariate Analysis–III. 1973. 383-407.

[21] De Jong, Sijmen. "SIMPLS: an alternative approach to partial least squares regression." Chemometrics and intelligent laboratory systems 18.3 (1993) 251-263.

[22] Martins, Joao Paulo A., Reinaldo F. Teofilo, and Márcia Ferreira. "Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets." Journal of Chemometrics 24.6 (2010) 320-332.

[23] Barker, Matthew, and William Rayens. "Partial least squares for discrimination." Journal of chemometrics 17.3 (2003) 166-173.

[24] Levine, Howard A. "AN Tikhonov and VY Arsenin, solutions of ill posed problems." Bulletin (New Series) of the American Mathematical Society 1.3 (1979) 521-524.

[25] Cui, Minshan, and Saurabh Prasad. "Class-dependent sparse representation classifier for robust hyperspectral image classification." IEEE Transactions on Geoscience and Remote Sensing 53.5 (2015) 2683-2695.

[26] Xu, Lu, et al. "Variable-weighted PLS." Chemometrics and intelligent laboratory systems 85.1 (2007) 140-143.

[27] Kim, Sanghong, et al. "Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection." International journal of pharmaceutics 421.2 (2011) 269-274.

[28] Song, Weiran, et al. "Differentiation of organic and non-organic apples using near infrared reflectance spectroscopy—a pattern recognition approach." SENSORS, 2016 IEEE. IEEE, 2016.

[29] Bache, Kevin, and Moshe Lichman. "UCI machine learning repository." (2013).

[30] Guyon, Isabelle, et al. "Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark." Pattern recognition letters 28.12 (2007) 1438-1444.

[31] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." Machine learning 53.1-2 (2003) 23-69.

[32] Tapp, Henri S., Marianne Defernez, and E. Katherine Kemsley. "FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils." Journal of agricultural and food chemistry 51.21 (2003) 6110-6115.

[33] Holland, J. K., E. K. Kemsley, and R. H. Wilson. "Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees." Journal of the Science of Food and Agriculture 76.2 (1998) 263-269.

[34] Zheng, Wenbin, Xiaping Fu, and Yibin Ying. "Spectroscopy-based food classification with extreme learning machine." Chemometrics and Intelligent Laboratory Systems 139 (2014) 42-47.

[35] Acquarelli, Jacopo, et al. "Convolutional neural networks for vibrational spectroscopic data analysis." Analytica chimica acta 954 (2017) 22-31.

[36] Smith, Benjamin R., Matthew J. Baker, and David S. Palmer. "PRFFECT: A versatile tool for spectroscopists." Chemometrics and Intelligent Laboratory Systems 172 (2018) 33-42.

[37] R.D. Snee, Validation of Regression Models: Methods and Examples, Technometrics. 19 (1977) 415–428.

[38] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, J.P. Huvenne, Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation, Chemom. Intell. Lab. Syst. 96 (2009) 27–33.

[39] Uchimaru, Taku, and Manabu Kano. "Sparse Sample Regression Based Just-In-Time Modeling (SSR-JIT): Beyond Locally Weighted Approach." IFAC-PapersOnLine 49.7 (2016) 502-507.

[40] E. Vigneau, P. Courcoux, R. Symoneaux, L. Guérin, A. Villière, Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception, Food Qual. Prefer. 68 (2018) 135–145.

[41] C. Strobl, J. Malley, G. Tutz, An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests, Psychol. Methods. 14 (2009) 323–348.