Chemometrics and Intelligent Laboratory Systems 154 (2016) 29–37

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Sparse regression for selecting fluorescence wavelengths for accurate prediction of food properties

Hiroshi Higashi[a,*], Gamal M. ElMasry[a,b], Shigeki Nakauchi[a]

[a] Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan
[b] Agricultural Engineering Department, Faculty of Agriculture, Suez Canal University, Ismailia, Egypt

ARTICLE INFO

Article history:
Received 21 July 2015
Received in revised form 7 March 2016
Accepted 8 March 2016
Available online 17 March 2016

Keywords:
Fluorescence
Excitation–emission matrix
Variable selection
Sparse regression

ABSTRACT

This paper tested various regression models (PLS, Ridge, Lasso, and sparse group Lasso) to select the appropriate fluorescence wavelengths/variables in excitation–emission matrices (EEMs) to improve the prediction of food identities. A framework using sparse models (the Lasso and sparse group Lasso) was proposed and compared with the conventional models. These sparse regression techniques can simultaneously achieve the ideal design of the estimator and select the most effective feature-related wavelengths. The experimental results showed that the proposed framework provided high prediction accuracy in selecting variables for accurate prediction of fish freshness and meat safety. Specifically, in case of predicting fish freshness, the sparse group Lasso regression had a determination coefficient R^2 of 0.790 with 493 EEM variables while the standard PLS regression had R^2 of 0.748 using all 1054 EEM variables.

1. Introduction

The excitation–emission matrix (EEM, also called a fluorescence fingerprint) is an arranged fluorescence spectrum of emission lights stimulated by lights with various excitation wavelengths [1]. Because the EEM implicates the physical and chemical properties of objects [1], it has been widely used for nondestructive measurement for various food properties [2–4]. Predicting a property of a target object with its EEM is considered to be a regression problem [5]. In such a problem, the explained variables are the fluorescence spectrum in the EEM of the target. The response variable is the desired property of the object to be predicted. Similar to usual regression problems, the real property and the corresponding EEM spectra of the training samples are measured to design an estimator (model) [5]. The model is designed in such a way that the model utilizes the explained variables to accurately predict the response variables of the samples [6]. In many cases, methods based on least squares such as the partial least squares (PLS) model [7] have been widely used. The EEMs of new samples (testing set) were used to predict their actual property by using the designed model.

In practice, the measurement of the EEMs provides a rapid quality monitoring of mass products [8,9]. To reduce the measurement time, the number of wavelengths of the excitation and emission lights should be reduced. However, the prediction with the less number of wavelengths could influence the accuracy of prediction in some circumstances [10]. A good variable selection technique cannot only capture variables that are most specifically related to the property on interest, but can also exclude regions affected by other sources of variation, leading to the enhancement of the model's robustness [11]. In reality, there is no standard method for wavelength selection because it is difficult to answer which algorithm (wavelength selection approach) is suitable for particular kind of data [12]. The choice of particular method depends on the nature of the problem, size of the data set, ease of implementation, and economic feasibility [13]. For instance, a method for wavelength selection based on grid-search has been proposed [10]. Because the method requires calculating the prediction accuracy for all wavelength candidates, the computational cost of the method is extremely large. Moreover, the number of the best wavelength candidates increases explosively as the size of the EEM increases. Therefore, development of reliable methods for rapid selection of wavelengths is a crucial issue to promote the use of nondestructive measurements using fluorescence spectra.

This paper proposes a new framework for selecting the best variables in EEM spectra. The discussed optimization problem can be considered to be a problem of finding sparse components in the EEMs related to the target property [14,15]. Therefore, models of sparse regression were applied to this problem. More specifically, the Lasso [14] and group Lasso [16,17] models were evaluated in the paper. The Lasso regression assumes that most of the weight coefficients for the explained variables become zero [15]. This means that measuring the fluorescence intensities corresponding to zero

[*] Corresponding author.

http://dx.doi.org/10.1016/j.chemolab.2016.03.008
0169-7439/© 2016 Elsevier B.V. All rights reserved.

coefficients is not needed to estimate the properties of new samples. Moreover, the group Lasso regression with sets of groups in which single excitation/emission wavelength belongs to the same group was proposed. The group Lasso works in such a way that all of the weight coefficients belonging to the same group are zero. This model can control the number of the groups, that is, the number of the excitation/emission wavelengths. Therefore, the estimator can be designed flexibly. For example, if an estimator as a hardware device [12,18,19] was designed, the group Lasso can control the number of filters needed to generate the excitation lights with narrow bandwidths taking economical cost and the number of the times for measurements into consideration. Additionally, the wavelength selection using the group Lasso can be efficiently applied to design the optical filters for a rapid measurement device [20]. The new feature brought by the proposed group set is called wavelength-wise variable selection in this paper.

The objectives of this paper were as follows:

- Developing a framework using sparse regression in the prediction problem using EEMs to provide efficient variable (wavelength) selection;
- testing the framework in food safety estimation as,
  - counts of viable bacteria on the surface of porcine meats, and
  - freshness condition of frozen fishes estimated by *K*-value; and
- discussing the efficiency of the sparse regression techniques in terms of the prediction accuracy, the number of the selected variables, efficiency of parameter tuning, and flexibility in the design of the estimator.

2. Regression models

For the prediction of a certain property of the target objects using their corresponding EEM spectra, the following regression models were tested. Additionally, Section 2.7 introduces methods for the variable selection combined with a regression technique. In this study, the explained variables $(x_1, x_2, \ldots, x_N)$ are the excitation–emission wavelengths in EEM spectra at which the fluorescence intensities were registered and the response variable (y) is either the count of bacterial colonies in meat or *K*-value of fish freshness.

2.1. Formulation of linear regression

Linear regression [5] is a problem of predicting a response variable $y \in \mathbb{R}$ from explained variables $\{x_n \in \mathbb{R}\}_{n=1}^{N}$ by

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_N x_N = \boldsymbol{w}^\top \tilde{\boldsymbol{x}}, \tag{1}$$

where $\hat{y}$ is a predicted value of y, $\tilde{\boldsymbol{x}}$ is the vector of the explained variables defined by

$$\tilde{\boldsymbol{x}} = [1, x_1, x_2, \ldots, x_N]^\top \in \mathbb{R}^{N+1}, \tag{2}$$

and $\boldsymbol{w}$ is the vector of the weight coefficients defined by

$$\boldsymbol{w} = [w_0, w_1, \ldots, w_N]^\top \in \mathbb{R}^{N+1}. \tag{3}$$

In this problem, the weight coefficients $\{w_n\}_{n=0}^{N}$ that can accurately predict y are numerically determined.

2.2. Least squares regression

The least squares error is one of the loss functions to design the coefficients from observed samples [5]. Let $\{\boldsymbol{x}_m \in \mathbb{R}^N, y_m \in \mathbb{R}\}_{m=1}^{M}$ be the M pairs of samples of the explained variables and response variable. The least squares method finds the coefficients $\boldsymbol{w}$ that minimizes the cost function [21]:

$$J(\boldsymbol{w}) = \sum_{m=1}^{M} \left(y_m - \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_m\right)^2, \tag{4}$$

where $\tilde{\boldsymbol{x}}_m = [1, \boldsymbol{x}_m^\top]^\top$.

2.3. Ridge regression

Regularization for the least squares method has been proposed to prevent overfitting or to solve ill-posed problems [21,22]. The regularization for an optimization problem is to add a penalty term, which represents additional information such as smoothness or bounds of the norm of $\boldsymbol{w}$, to the cost function (Eq.(4)). Among regularization techniques being proposed, the Ridge regression [23] adds the l_2-norm of the weight vector as the penalty term. The l_2-norm for a vector is denoted as $\|\cdot\|_2$ and defined as $\|\boldsymbol{a}\|_2 = (|a_1|^2 + \cdots + |a_N|^2)^{\frac{1}{2}}$, where $\boldsymbol{a}$ is a vector defined as $\boldsymbol{a} = [a_1, \ldots, a_N]^\top$. It is defined as

$$\min_{\boldsymbol{w}} \ \frac{1}{2} J(\boldsymbol{w}) + \beta \| \boldsymbol{w} \|_2^2, \tag{5}$$

where $\beta \in \mathbb{R}^+$ is the regularization coefficient adjusting the effect of the penalty term.

2.4. Lasso regression

Instead of l_2-norm used in the Ridge regression, the Lasso regression [14] uses the l_1-norm of the weight vector as the penalty term. The l_1-norm for a vector is denoted as $\|\cdot\|_1$ and defined as $\|\boldsymbol{a}\|_1 = |a_1| + \cdots + |a_N|$. The optimization problem for Lasso regression is defined as

$$\min_{\boldsymbol{w}} \ \frac{1}{2} J(\boldsymbol{w}) + \alpha \| \boldsymbol{w} \|_1, \tag{6}$$

where $\alpha \in \mathbb{R}^+$ is the regularization coefficient. The Lasso regression promotes the sparsity of the coefficient vector by minimizing the l_1-norm of the weight vector [24]. Optimizing the problem in Eq. (6) needs iterative methods such as linear programming [25] or gradient methods [26–28]. In this paper, optimizing the weight coefficients was performed by applying the coordinate descent algorithm [29].

2.5. Group Lasso and sparse group Lasso regressions

The group Lasso regression [16] adds the sum of the l_2-norm in groups of the coefficients in the weight vector as the penalty term. The optimization problem for the group Lasso is defined as

$$\min_{\boldsymbol{w}} \ \frac{1}{2} J(\boldsymbol{w}) + \gamma \sum_{q=1}^{Q} \sum_{g \in G_q} w_g^2, \tag{7}$$

where $\gamma \in \mathbb{R}^+$ is the regularization coefficient, Q is the number of the groups, and G_q is the sets of the indexes of the weight coefficients in the qth group.