Chemometrics and Intelligent Laboratory Systems 153 (2016) 33–39

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Selecting variables with the least correlation based on physarum network

Tong Chen [a,c,*,1], Xing-Cong Zhao [a,b,1], Hang Zhou [a,c], Guang-Yuan Liu [a,c,*]

[a] School of Electronic and Information Engineering, Southwest University, Chongqing 400715, P. R. China
[b] Chongqing Academy of Metrology and Quality Inspection, 401121 Chongqing, P. R. China
[c] Chongqing Key Laboratory of Nonlinear Circuit and Intelligent Information Processing, Southwest University, Chongqing 400715, P. R. China

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 5 February 2016
Accepted 16 February 2016
Available online 24 February 2016

Keywords:
Variable selection
Wavelength selection
Real-time spectroscopy
Physarum network

ABSTRACT

In spectroscopy, redundant information makes the number of input variables for a prediction model larger than required. We present a method based on the physarum network to select the variable with the least correlation. This method transforms the variable selection problem into a path finding problem and then solves the problem based on the mechanism of foraging of *Physarum polycephalum*. Experimental results show that the physarum network, combined with other feature selection or extraction methods, can select the least number of wavelengths without sacrificing the prediction performance.

1. Introduction

Spectroscopy has been widely used to quantitatively analyze complex samples because of its noninvasive, effective and speedy manner. To achieve the spectroscopy analysis, a prediction model has to be developed. It relates the spectral variables (spectra of samples) and the property values (e.g., chemical concentration) of the samples. However, the spectral variables are normally high dimensional data, i.e., the number of spectral variables is large, and redundant information is contained in the variables. The large number of spectral variables can not only make the prediction unreliable but also complicate the prediction model and calculation. Therefore, dimensionality reduction, which aims to find the minimum number of variables (intrinsic dimensionality) necessary to explain the property values [1], is usually employed before the determination of a model.

The methods used to reduce dimensionality can be grouped into two types, i.e., feature extraction and feature selection from the point of view of machine learning or statistics [2]. The feature extraction methods project data from a high dimensional space into a low dimensional space, e.g., principle component analysis/regression (PCA/PCR) [3], partial least square regression (PLS) [4], etc. The feature selection methods find a subset of the original data, e.g., genetic algorithm (GA) [5], interval partial least squares regression (iPLS), etc. [6].

The feature extraction and feature selection methods can be used together to further reduce the dimensionality without sacrificing the prediction accuracy. A good example of this combination is GA-PLS. Because the PLS considers the information of both independent variables and dependent variables to extract principle components [4], researchers initially thought that PLS performed over the whole spectrum was good enough to reduce the dimensionality. However, it was later found that PLS used together with GA can both simplify the prediction model and improve the prediction accuracy [7]. Thus, GA-PLS has become one of the most popular variable selection methods [8–11].

Several variable selection methods regard the feature selection problem as an optimization problem. However, variable selection can also be performed based on prior knowledge, such as peak absorbance of the target components [6]. Though this selection method is effective and can be used together with other dimensionality reduction methods, it is less popular, probably due to the difficulty in obtaining or understanding the prior knowledge [6]. In this paper, we present a variable selection method based on the characteristics of spectroscopy sensors. The proposed method can be used together with optimization-based variable selection or feature extraction methods to further reduce dimensionality without sacrificing prediction accuracy, i.e., it uses the minimum number of variables to predict property values. This characteristic of the proposed method is important because it reduces the calculation time and simplifies the prediction model, thus making the method potentially suitable for real-time spectroscopy application. The prior knowledge needed includes the spectral bandwidth and increasing step of the spectrograph (see Section 2.1), which are easy to understand and obtain.

Specifically, we use a physarum network (PN) [12] to search the spectrum to obtain a subset of spectral variables with the least correlation. The PN is used before GA or after iPLS. Experimental results show that the PN-GA-PLS or iPLS-PN-PLS method can predict property values

* Corresponding authors at: School of Electronic and Information Engineering, Southwest University, Chongqing 400715, P. R. China. Tel.: +86 23 68252520.
E-mail addresses: c_tong@swu.edu.cn (T. Chen), liugy@swu.edu.cn (G.-Y. Liu).
[1] These authors contribute equally to this work.

http://dx.doi.org/10.1016/j.chemolab.2016.02.007
0169-7439/

with a similar performance compared to GA-PLS or iPLS but with fewer spectral variables.

2. Theory

2.1. Spectrograph and redundant spectral information

A spectrograph is a key part of spectroscopy instrumentation. It splits the whole spectrum of white light into sub-spectral bands. The reflectance or absorbance (spectral variable) of a sample at each sub-spectral band can then be measured and related to the property values of the sample. There are three common categories of spectrographs [13], i.e., the dispersive spectrograph, the Fourier transform interferometer and the narrow-band tunable filter. The sub-spectral bandwidth is determined by the setting of the spectrograph, e.g., the slit of the dispersive spectrograph determines the sub-spectral bandwidth.

However, a spectrograph can normally measure the spectral response with a step smaller than the sub-spectral bandwidth. For example, a dispersive spectrograph with a 30 μm slit can give a 5 nm bandwidth but can measure the spectral response with a 2 nm step (resolution). This relationship between bandwidth and wavelength increasing step is illustrated in Fig. 1. It is obvious that each spectral variable at every spectral band includes information regarding the spectral variables of the adjacent spectral band. Within a spectral range including several spectral variables, one or two spectral variable(s) would be enough to carry all of the information necessary for predicting property values.

Random correlation between spectral variables can also lead to redundant spectral information in selected variables. When a ratio of variables to samples is equal to or larger than 5, the random correlation will be so severe that it would be dangerous to use GA [7]. A method for avoiding this random correlation could be creating new variables by averaging the original ones [7]. In this way, the ratio variables/sample could decrease to 5. However, creating one new variable by averaging several adjacent spectral variables is a method that will reduce the spectral resolution. This averaging method may be opposite to the core idea of using a spectrograph or spectroscopy, i.e., to measure radiation intensity on narrower wavebands. In some cases, a property value may only be related to a very narrow waveband with one to two spectral variables. Averaging these variables would not increase the final prediction performance. An alternative way for solving the random correlation problems would be to select variables with the least correlation.

2.2. Physarum network (PN) and its mathematical model

Physarum polycephalum is a slime mold. It can be in vegetative phase that is called plasmodium. The plasmodium is an amoeba-like organism with a body shape of a dendritic network consisting of tubular components. Nakagaki et al. [14] conducted an interesting experiment in 2000. They put plasmodium in a maze with two food sources: one at the entrance and the other at the exit of the maze. It was found that the plasmodium changed its body shape to connect the two food sources (entrance and exit); moreover, the plasmodium always connected the two points using the shortest length of tubes, i.e., it finds the shortest route in the maze.

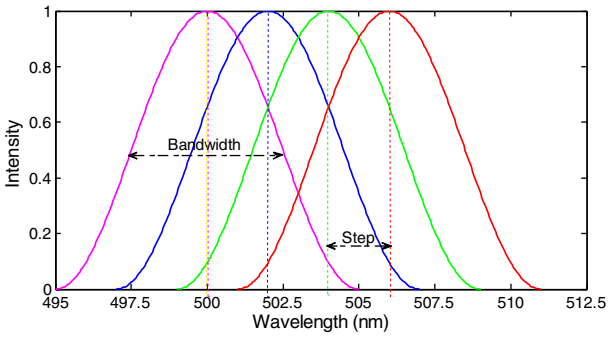

Fig. 1. Illustration of bandwidth (5 nm) and measuring step of a spectrograph (2 nm).

After investigating the physiological background of plasmodium growth, Tero et al. [15,16] developed a mathematical model based on the physarum network for path finding. This model was proven to be able to find the shortest route by Bonifaci [12] through mathematical deduction. By advancing the physarum model, Liu et al. [17,18] developed a physarum optimization algorithm, which is suitable for solving the Steiner tree problem with low complexity and high parallelism.

Tero's model assumes that the flowing of nutrients inside the tubes of the physarum network is driven by the pressure due to the rhythmic contractions of the tubes, with the entrance node of the maze being the source of the nutrient flow and the exit node being the sink of the nutrient flow.

The nutrient flux (nutrient flow per unit area) through node i (N_i) to node j (N_j) in the maze is expressed as Q_{ij}, which can be computed by using the formula

$$Q_{ij} = \frac{\pi r_{ij}^4 (P_i - P_j)}{8\eta L_{ij}}, \tag{1}$$

where P_i is the pressure at the node N_i, η is the viscosity coefficient of the flow, and r_{ij} and L_{ij} are the radius and length, respectively, of the tube connecting N_i to N_j. If there is more than one tube connecting N_i and N_j, then Q_{ij} (r_{ij}, L_{ij}) can be $Q_{ij\_1}(r_{ij\_1}, L_{ij\_1})$, $Q_{ij\_2}(r_{ij\_2}, L_{ij\_2})$, or $Q_{ij\_3}(r_{ij\_3}, L_{ij\_3})$, representing the first, second and third tubes, respectively.

A variable $D_{ij} = \frac{\pi r_{ij}^4}{8\eta}$, i.e., the conductivity representing the ability to conduct the flow, is defined in the model; thus, Eq. (1) can be rewritten as

$$Q_{ij} = \frac{D_{ij}(P_i - P_j)}{L_{ij}}. \tag{2}$$

Except for source (N1) and sink node (N2), each node is assumed to be zero capacity. According to the conservation law of flow, the sum of flux at each node can be

$$\sum_i Q_{ij} = 0, j \neq 1, 2. \tag{3}$$

For the nodes N1 and N2, the flux equations are

$$\sum_j Q_{1j} - I_0 = 0 \tag{4}$$

and

$$\sum_j Q_{2j} + I_0 = 0, \tag{5}$$

where I_0 is the flux from the source node to the sink, which is assumed to be constant in the model.

According to Eqs. (3)-(5), Eq. (2) can then be rewritten as

$$\sum \frac{D_{ij}}{L_{ij}}(P_i - P_j) = \begin{cases} I_0, & j = 1 \\ -I_0, & j = 2 \\ 0, & otherwise \end{cases}. \tag{6}$$

The conductivity D_{ij} is assumed to change when adapting to the flux Q_{ij}. Moreover, the tubes with zero conductivity will die out. The evolution of D_{ij} is expressed as

$$dD_{ij}/dt = f(|Q_{ij}|) - D_{ij}. \tag{7}$$