The application of R language in the selection of characteristic bands for the prediction of protein content in milk powder by Near Infrared Spectroscopy

Huixin Guo1*, Yaolan Yin2, Fang Zheng2 and Pingzhen Li2

1College of Food Science and Nutritional Engineering, China Agricultural University, Beijing, 100083, China
2College of Information, Shanxi University of Finance and Economic, Taiyuan, Shanxi, 030006, China

*Corresponding author’s e-mail: guohuixin@cau.edu.cn

Abstract. Milk powder is an important food, suitable for preservation and transportation. Protein is an important nutritional component in milk powder. At present, the routine physical and chemical analysis method is usually used in the detection of components of milk powder, which is time-consuming and labor-consuming. Therefore, it is very important to carry out the rapid non-destructive detection of milk powder quality. The application of near infrared spectroscopy (NIRS) technology in the rapid detection of milk powder quality is increasingly mature, but many analytical techniques are not perfect. In this study, the models of prediction protein in milk powder were established by R language with NIRS, and the characteristic bands were selected by recursive feature extraction (RFE) in R language, and the selected bands were screened one by one to determine their importance. Finally, using Partial least squares (PLS), generalize linear model (GLM), support vector machine (SVM), least angle regression (LARS), linear model (LM) and other methods to build the prediction models with 8 characteristic bands, the R-squared of the models can be increased by 0.1 to 0.7, the model robustness greatly improved.

1. Introduction

Milk powder is a kind of powder made by removing almost all water from animal milk and adding proper amount of vitamins and minerals. It is suitable for preservation and convenient for transportation. The production of milk powder originated in France of 1800s. At present, most milk powder is made by spray method invented by Americans in the 1870s[1]. Protein is an important nutritional component of milk powder. Protein is also an important component of all cells and tissues in human body. Protein is an organic macromolecule, which must contain carbon, hydrogen, oxygen and nitrogen. The routine method for the determination of physical and chemical indexes of milk powder is time-consuming and tedious, which cannot meet the requirements of online analysis in modern industry. How to use a more efficient and accurate detection method is an important work.

Near infrared spectrum refers to electromagnetic wave with wavelength in 780-2526nm, which is between visible spectrum and mid infrared spectrum. The frequency doubling and combined absorption of the vibration of the hydrogen group X-H (X= C, N, O) are recorded, which is very suitable for the qualitative and quantitative measurement of organic substances containing hydrocarbons. NIRS is characterized by rapidity, no need of reagents, safety and high efficiency.
NIRS has been widely used in the field of food since 1970s and milk powder detection since 1990s[2]. However, at present, many spectral analysis software is usually simple, and the modeling method is single, such as TQ software. There is only one modeling method for each type of model. The characteristic bands selection method is also very single and cannot set the required parameters. It can only basically meet the industrial application, but cannot carry out scientific research in a deeper level.

In this study, programs independently written in the R language were used for preprocessing of the spectrum, model establishment, selection of characteristic bands, and model verification. The emphasis is selection of characteristic bands in this study. Through the selection of characteristic bands by improved RFE method, we can get the better model than before.

2. Materials and methods

2.1. Materials
The experimental samples come from the markets all over the country. Different brands and different kinds of milk powder samples are collected. The total number of samples is 150.

2.2. Spectrum acquisition
The diffuse reflectance spectra of milk powder samples were collected by using the integrating sphere accessory of antaris Fourier transform near infrared spectrometer (Thermonicolet, USA). In order to eliminate the spectral change caused by the difference of sample loading, the auto rotating sample cup accessory is adopted, the spectral range is 4000-10000 cm\(^{-1}\) (1000-2500 nm), the resolution is 8 cm\(^{-1}\), and the average value is set for 32 scans each time. The height of each sample is 2 cm, each sample is collected three times, and the average spectrum of three spectra is taken as the final sample spectrum.

2.3. Determination of chemical value
The protein content of milk powder was determined by Kjeldahl method[3]. The average value of each sample for three times was taken as the protein content of the final sample. Table 1 shows the statistical results of the protein content of the samples in the calibration and prediction set. It can be seen that the samples have good representativeness and wide distribution range. The sample information in the prediction set is well covered by the modeling set.

|                  | Number | Average | Max | Min | SD   |
|------------------|--------|---------|-----|-----|------|
| Total            | 150    | 15.56   | 24.3| 10.7| 3.9384|
| Prediction set   | 37     | 15.68   | 21.1| 11.9| 2.7460|
| Calibration set  | 113    | 15.52   | 24.3| 10.7| 2.6462|

2.4. Data processing and analysis software
In this study, R language is used for all analysis. Near infrared spectrum data is high-dimensional, and contains a lot of noise information, which is not only useless for analysis, but also causes certain interference to the data of target substances, so the preprocessing method is very important. In this study, the improved multiple scattering correction (MSC) and the first derivative method are used for preprocessing. PLS, GLM, SVM, LARS, LM and other methods are used for modeling comparison. RFE are used to select characteristic bands. R language is an operating environment commonly used in statistical analysis and mapping. It can realize a variety of statistical analysis methods, and update much faster than general statistical software such as SPSS and SAS. At the same time, in R language, each method can set its own parameters which are most suitable for analysis and get better results.
3. Results and discussion

3.1. Preprocessing of spectral data

The original spectrum of the milk powder sample is shown in Figure 1. MSC and first derivative can effectively eliminate baseline drift caused by particle size and particle inequality. Due to the difference in the density and granularity of the milk powder during loading, the spectral baseline of the milk powder sample is shifted. The original spectrum of the milk powder is processed by MSC and first derivative in R language to eliminate the baseline drift of the light. The processed spectra were as follows shown in Figure 2.

![Figure 1. Original spectra of raw milk powder samples.](image1)

![Figure 2. Spectra of processed by MSC and First derivative.](image2)
3.2. Preprocessing of spectral data

The spectrum collection range of milk powder samples is 1000-2500 nm, and there are 1557 wavelength points. PLS method[4] is always a good modeling method when establishing the model of near-infrared spectrum analysis. However, in different samples and environments, other methods often show better results than PLS. In this study, different modeling methods such as PLS, GLM[5], SVM, LARS and LM were used to build protein model with 1557 wavelengths. The modeling results are shown in Table 1. The R-squared of the model established by PLS is 0.8733, and is the best. In this study, the R-squared of GLM protein prediction model is 0.8682, which is close to that of PLS.

### Table 2. Prediction models established by original spectrum.

| Method     | RMSE     | R-squared | MAE     |
|------------|----------|-----------|---------|
| PLS        | 0.9734   | 0.8733    | 0.7156  |
| GLM        | 1.0015   | 0.8682    | 0.7454  |
| SVM (Linear) | 1.3720   | 0.7685    | 1.0797  |
| SVM (Radial) | 1.1264   | 0.8271    | 0.8040  |
| LARS       | 2726.1   | 0.7416    | 1961.5  |
| LM         | 136.73   | 0.1470    | 108.62  |

3.3. Selection of characteristic bands

The purpose of selection of characteristic bands is to reduce the number and dimension of features, enhance the generalization ability of the model, reduce over fitting, and enhance the understanding between indicators and characteristic bands. The main work of this study is to select a subset of 1557 bands to build a better prediction model without reducing R-squared.

3.3.1. Selecting characteristic bands with RFE. The main idea of recursive characteristic elimination is to traverse all characteristic, build the model repeatedly, and then select the best characteristic. “RFFuncs” is used as a function to build the model repeatedly in R language. The random forest method used by “RFFuncs” function is an algorithm that uses multiple decision trees to train samples. It can be used to calculate the importance of characteristic and remove irrelevant characteristic according to the importance of characteristic. The relationship between the number of bands displayed in the RFE program and RMSE is shown in Figure 3. When there are 87 bands in the model, RMSE is the minimum. The positions of 87 selected characteristic bands on the full spectrum are shown in Figure 4. Among them, 26 wavelength points are selected in the range of 1004.2-1197 nm, 22 in the range of 1490.1-1760.2 nm, and 39 in the range of 1937.8-2500 nm.
Six methods are used to built models with 87 selected wavelengths, and the results are shown in Table 2. From the results, the models built by SVM, LARS and LM methods have been greatly improved.

Table 3. Models based on the selected characteristics bands by RFE.

| Method         | RMSE  | R-squared | MAE  |
|----------------|-------|-----------|------|
| PLS            | 1.0222| 0.8551    | 0.7608|
| GLM            | 1.0420| 0.8501    | 0.7778|
| SVM (Linear)   | 1.0071| 0.8556    | 0.7450|
| SVM (Radial)   | 1.0963| 0.8312    | 0.7760|
| LARS           | 1.3352| 0.7811    | 1.0365|
| LM             | 2.1871| 0.5483    | 1.7765|

3.3.2. Selecting less characteristic bands with improved RFE method. Among the characteristics bands selected by RFE method, there are also redundant features. On this basis, this study designs a new scheme to remove redundant bands, that is, programming in R language, screening all bands one by one, and identifying whether they are characteristics bands that affect the performance of the model. Through identification, it is finally simplified to keep only 8 characteristics bands. The positions of these 8 characteristics bands in the full spectrum are shown in Figure 5. They are 1013 nm, 1007 nm, 1664 nm, 1748 nm, 1758 nm, 2066 nm, 2073 nm and 2106 nm respectively.

![Figure 5. Location of characteristic bands after the improved method by RFE.](image)

The models are established by using the 8 feature bands after the final reduction, and the model results of the 6 methods are shown in Table 3. From the results, improved RFE method can effectively improve the performance of models built by SVM, LARS and LM methods. In particular, the model built by LM method has been basically available since it was very poor at the beginning.

Table 4. Models based on the selected characteristics bands by improved RFE method.

| Method    | RMSE  | R-squared | MAE  |
|-----------|-------|-----------|------|
| PLS       | 1.0110| 0.8536    | 0.7740|
| GLM       | 0.9973| 0.8591    | 0.7573|
| SVM (Linear) | 0.9881| 0.8604    | 0.7520|
4. Conclusion
At present, the development of NIRS technology is in a bottleneck period, and its application in many fields has been very perfect, such as food, textile, petroleum, materials and cultural relics identification, etc., but the perfect theoretical support and new and more application fields have made slow progress. This is related to the development of near-infrared analysis methods. The last time near-infrared technology developed rapidly, it was the era of widespread computer technology. How to break through the bottleneck of near infrared technology depends on computer technology. One of the key technologies in near infrared technology is to select characteristics bands. Choosing the right characteristics bands can improve the performance of the model. The most important thing to make a model is whether the characteristics bands has a relationship with the target material. If there is a relationship, a good model can be made even if there are few characteristics bands. If there is no relationship, a good model cannot be made if there are more characteristics bands. Therefore, it is very important to extract characteristics bands, while the existing software such as TQ and OPUS has a single method to select characteristics bands, and they cannot set relevant parameters, and cannot accurately select features suitable for analysis. It is a good method to select the appropriate characteristic variable by programming in R language environment and adjusting the parameters suitable for the analysis environment.

References
[1] Koc, A.B., Heinemann, P.H., & Ziegler, G.R. (2003) A process for increasing the free fat content of spray-dried whole milk powder. J. Food Sci., 68(1): 210-216.
[2] Inácio, M.R.C., de Moura, M. de F.V., & de Lima, K.M.G. (2011) Classification and determination of total protein in milk powder using near infrared reflectance spectrometry and the successive projections algorithm for variable selection. Vib. Spectrosc., 57(2): 342–345.
[3] Kamizake, N.K.K., Goncalves, M.M., Zaia, C.T.B.V., & Zaia, D.A.M. (2003) Determination of total proteins in cow milk powder samples: a comparative study between the kjeldahl method and spectrophotometric methods. J. Food Compos. Anal., 16(4): 507-516.
[4] Lim, J., Kim, G., Mo, C., Kim, et al. (2016) Detection of melamine in milk powders using near-infrared hyperspectral imaging combined with regression coefficient of partial least square regression model. Talanta, 151: 183–191.
[5] Hastie, T. J. (2015). The elements of statistical learning. Springer Publishing, Berlin.