The NIR Corn Data Set

Hongwei PENG

Supervisor : Prof Tom Fearn

Department of Statistical Science
University College London

August 4, 2019

Contents

The most readily available high-dimensional NIR spectroscopic data is called corn data. There are many algorithms for analysing corn data in many publications. These algorithms will often claim that their new algorithm has a better performs. So the purpose of this dissertation is to search for as many different papers as possible, and write a critical overall to find the most efficient measure that can evaluate whether the model performs well and quantify the improvements mentioned in the paper.

1 Introduction

2 Literature reviews

3 Datasets

4 Methodology

4.1 Model Evaluation

According to the corn data literature, there are several measures that can be used to evaluate the performance of the model.

1, Root Mean Square Error for Calibration samples (RMSEC) is proposed by Yun et al. (2014).

2, RMSECV is mentioned by many papers (Ji et al., 2015). And there are two cross-validation methods. One is leave one out cross-validation (LOOCV), mentioned by Zheng et al. (2015). The other is K-fold cross-validation, there are the 3-fold cross-validation (Galvão et al., 2007), 10-fold cross-validation (Ji et al., 2015) and so on. The calculation method of RMSECV is as follows:

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

where: n = the number of samples
y_i = the experimental value of the i-th sample
$\hat{y}_i$ = the predicted value of the i-th sample by cross-validation which includes removing the set of i-th sample from the calibration set, building a model with the remaining samples, and applying the model to i-th sample

3, The Root Mean Square Error of Prediction (RMSEP) is mentioned by Su et al. (2006). This is a generally accepted method of evaluating models. This approach requires the determination of appropriate cross-validation sets and prediction sets before building PLS model. For example, 60 samples of corn data are used for a cross-validation and the remaining 20 samples are used as predictions (Su et al., 2006). Then the cross-validation data is used for modelling, determining the parameters for regression model, such as PLS. After that, the model is applied to the predictive data to calculate the Root Mean Square Error of Prediction (RMSEP). The RMSEP calculation formula is:

$$RMSEP = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{4.2}$$

where: m = the number of prediction sets
y_i = the experimental value of the i-th sample in the prediction set
$\hat{y}_i$ = the prediction value of model for the i-th sample

4, There are few papers mentioned that R^2 is used to measure the model (Tatavarti et al., 2005). But this method is also flawed. In some situations, not enough calculation accuracy of the computer will cause the value of R^2 to be equal to 1. For example, Deng et al. (2016) has this problem and the R^2 in PLS model fitting moisture is equl to 0.9959, but for the CARS, GA-PLS and BOSS model, R^2 are all equal to 1.0000 ± 0.0000. Hence we can see it hard to distinguish the different between models. So this will not be a good indicator of evaluation.

4.2 Pre-treatment

The papers use different pre-treatments of the data, and the results of the model will be very different. For example, 7th paper and 8th paper also use M5 to predict the first constituent, and the results are very different. Through the different literatures, the common pretreatment of corn data has the following four types:

1. Nothing to deal with, such as 1st paper.

2. Scale the data, such as 4th paper.

3. SavitzkyGolay filter processing on the data, such as 3rd paper.

4. Delete the outliers, such as 8th paper.

4.3 Sample splitting

The number of samples selected and the method of sample selection will also have an impact on the results of models. The methods of selecting samples are as follows:

1. A completely random sample (Su et al., 2006).

2. Choosing the samples by the SPXY method (Galvão et al., 2007).

3. Use the Kennard-Stone (K-S) algorithm, such as 8th paper in the third form.

4. Directly divide the raw data into the first half and the second half, the first half is cross-validation sets, and the second half is prediction set. For example 4th paper.

The second and third methods will result in the prediction level data being very close to the data of the cross-check set, which may increase the accuracy of the prediction and reduce the prediction difficulty, so these two methods are not used here. Because in the actual problem, the performance of the algorithm looks better when the two sample sets are too similar, which is not what we want. The last method relies on the sorting of the original data, so the next step is to take a random sampling method to simulate the sample. The selection of the quantity

refers to the 5th and 6th documents, so the sample size of 20 70 will be selected to build the model, and the rest will be predicted.

4.4 Cross-validation

4.5 PLS algorithm

4.6 Parallel computing

4.7 Myriad

5 Result and discussion

5.1 Loop times

5.2 Number of Samples

5.3 Number of Components

5.4 Pre-treatment

5.5 Cross-Validation

5.6 Compare with papers

6 Conclusions

7 Conclusion

References

Deng, B.-C., Y.-H. Yun, D.-S. Cao, Y.-L. Yin, W.-T. Wang, H.-M. Lu, Q.-Y. Luo, and Y.-Z. Liang (2016). A bootstrapping soft shrinkage approach for variable selection in chemical modeling. *Analytica chimica acta 908*, 63–74.

Galvão, R. K. H., M. C. U. Araújo, E. C. Silva, G. E. José, S. F. C. Soares, and H. M. Paiva (2007). Cross-validation for the selection of spectral variables using the successive projections algorithm. *Journal of the Brazilian Chemical Society 18*(8), 1580–1584.

Ji, G., G. Huang, Z. Yang, X. Wu, X. Chen, and M. Yuan (2015). Using consensus interval partial least square in near infrared spectra analysis. *Chemometrics and Intelligent Laboratory Systems 144*, 56–62.

Su, Z., W. Tong, L. Shi, X. Shao, and W. Cai (2006). A partial least squares-based consensus regression method for the analysis of near-infrared complex spectral data of plant samples. *Analytical letters 39*(9), 2073–2083.

Tatavarti, A. S., R. Fahmy, H. Wu, A. S. Hussain, W. Marnane, D. Bensley, G. Hollenbeck, and S. W. Hoag (2005). Assessment of nir spectroscopy for nondestructive analysis of physical and chemical attributes of sulfamethazine bolus dosage forms. *aaps Pharmscitech 6*(1), E91–E99.

Yun, Y.-H., W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, and Q.-S. Xu (2014). A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Analytica chimica acta 807*, 36–43.

Zheng, K.-Y., X. Zhang, P.-J. Tong, Y. Yao, and Y.-P. Du (2015). Pretreating near infrared spectra with fractional order savitzky–golay differentiation (fosgd). *Chinese Chemical Letters 26*(3), 293–296.