A new regularised discriminant analysis; principal discriminant variate method for handling multicollinear data

Jian-Hui Jiang,[a,b] Roumiana Tsenkova,[c] Ru-Qin Yu[b] and Yukihiro Ozaki[a*]

[a]*Department of Chemistry, School of Science and Technology, Kwansei-Gakuin University, Sanda, Hyogo 669-1337, Japan*

[b]*College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China*

[c]*Department of Environmental Information and Bioproduction Engineering, Faculty of Agriculture, Kobe University, Noda-ku, Kobe 657-8501, Japan*

Introduction

In linear discriminant analysis there are two important properties concerning the effectiveness of discriminant function modelling.[1–10] The first is the separability of the discriminant function for different classes. The separability reaches its optimum by maximising the ratio of between-class to within-class variance. The second is the stability of the discriminant function against noise present in the measurement variables. One can optimise the stability by exploring the discriminant variate in a principal variation subspace, i.e. the directions that account for a majority of the total variation of the data. An unstable discriminant function will exhibit inflated variance in the prediction of future unclassified objects and be exposed to a significantly increased risk of erroneous prediction. Therefore, an ideal discriminant function should not only separate different classes with a minimum misclassification rate for the training set but also possess a good stability so that the prediction variance for unclassified objects can be as small as possible. In other words, an optimal classifier should find a balance between the separability and the stability. This is of special significance for multivariate spectroscopy-based classification where multicollinearity always leads to discriminant directions located in low-spread subspaces.

We have developed, recently, a new regularised discriminant analysis technique, the principal discriminant variate (PDV) method, which handles effectively multicollinear data commonly encountered in multivariate spectroscopy-based classification.[11] The motivation behind this method is to seek a sequence of discriminant directions that not only optimise the separability between different classes but also account for a maximised variation present in the data. Three different formulations for the PDV method have been suggested and an effective computing procedure has been proposed for PDV.[11] The PDV method has been applied to near infrared (NIR) spectra of whole blood samples from mastitic and healthy cows to demonstrate the potential of the PDV method in clinical diagnosis of mas-

Near Infrared Spectroscopy: Proceedings of the 10th International Conference

titis.[11] Mastitis is an udder inflammation that is a major problem for the global dairy industry.[12] It causes substantial economic losses by decreasing milk production, changing the compositions in milk considerably, reducing milk quality and increasing the risk of early culling of cows. The NIR spectra of the whole blood samples from the mastitic and healthy cows have been clearly discriminated by the PDV method.

Experimental

The preparation of cow blood samples and the instrumentation and experimental procedure for measuring their NIR spectra were described in Reference 11. A total of 162 whole blood samples were prepared and separated into two classes, one associated with the cows suffering from mastitis and the other corresponding to cows not suffering from mastitis. These samples were divided into two sets, the training set and the prediction set. The training set contained 80 samples from Class 1 and 42 samples from Class 2, respectively. The prediction set was composed of the remaining samples, i.e. 25 samples from Class 1 and 15 samples from Class 2, respectively. The algorithm and software used for principal component analysis (PCA), discriminant partial least squares (DPLS), Fisher linear discriminant analysis (FLDA) and soft independent modeling of class analogies (SIMCA) were reported in Reference 11.

Theory and methods

Fisher linear discriminant analysis

In FLDA, one finds a linear combination of the original variables that would maximise the within-class variability relative to the between-class variability.[1,2] Suppose there are N objects $\mathbf{x}_n$ (n = 1, 2, —-, N) from K classes $\mathbf{C}_k$, (k = 1, 2, —-, K) with the kth class containing N_k objects, then the between-class, within-class and total covariance matrices are represented as follows:

$$\mathbf{B} = \frac{1}{N}\sum_{k=1}^{K} Nk(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathrm{T}} \tag{1}$$

$$\mathbf{W} = \frac{1}{N}\sum_{k=1}^{K}\sum_{\mathbf{x}_i\in\mathbf{C}_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathrm{m}_k)^{\mathrm{T}} \tag{2}$$

$$\mathbf{T} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}} \tag{3}$$

$$(\mathbf{T} = \mathbf{B} + \mathbf{W})$$

Here, $\mathbf{m}$ and $\mathbf{m}_k$ are the mean vector of all objects and those from $\mathbf{C}_k$, respectively, as given by

$$\mathbf{m} = \frac{1}{N}\sum_{n=1_k}^{N} \mathbf{x}_n \tag{4}$$

$$\mathbf{m}_k = \frac{1}{N}\sum_{\mathbf{x}_i\in\mathbf{C}_k} \mathbf{x}_i \tag{5}$$

In FLDA, one must maximise over **a** the following formulation;

$$J = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a}} \tag{6}$$

or equivalently

$$J = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{T}\mathbf{a}} \tag{7}$$

The drawback of FLDA is that, in multicollinear situations, it is inclined toward overfitting the training data. This would generally result in discriminant variates located in a small-variance subspace, hence showing instability or inflated variance in the prediction of unclassified objects. FLDA maximises the separability without regard to the stability of the classifier, since its formulation does not take account of the variance explained by the discriminant variates. In contrast, PCA aims at seeking the components that represent as much total variation of the data as possible, ignoring the separability of different classes. Therefore, PCA sacrifices the separability to approach maximum stability. Now, one may ask if an integration of PCA and FLDA can result in a discriminant analysis method furnished with favourable separability along with improved stability. This is the motivation behind the proposed PDV method.

Principal discriminant variate method

The PDV method is intended to seek a compromise between PCA and FLDA such that the stability of the method can be improved substantially without significant loss of discriminatory capability. In PDV, one must find a direction, called principal discriminant variate, which maximises the following principal discrimination criterion;

$$J = \frac{\mathbf{a}^{\mathrm{T}}\left[(\lambda\mathbf{B} + (1 = \lambda)\mathbf{T}\right]\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\left[\lambda\mathbf{T} + (1 - \lambda\mathbf{I}\right]\mathbf{a}} \tag{8}$$

where **I** is the identity (unitary) matrix of the same size as the covariance matrix **T** and λ, with a value varied between 0 and 1, is a weight controlling the balance between linear discriminant analysis (LDA) and PCA. By setting $\lambda = 1$, one immediately finds that the principal discrimination criterion (8) is reduced to Eq. (6), the discrimination criterion used in ordinary FLDA. On the other hand, if one sets $\lambda = 0$, the principal discrimination criterion (8) becomes

$$J = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{T}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{a}} \tag{9}$$

which turns out to be the variance criterion maximised in PCA.

If one increases the value of λ, the separability is increased. On the other hand, if the value of λ is decreased, the stability is increased. Thus, λ controls the trade-off between separability and stability.

The optimal value of λ can be taken as a value that yields a minimum error rate found by cross-validation or extra set of test objects. In practical applications, the performance of the PDV method is not so sensitive to the choice of the value of λ so that it is enough to optimise the parameter stepwise. That is, λ can be started from a small value, say 10^{-7}, then increased stepwise by a factor of about 10. For each value of λ one evaluates the misclassification rate on the training set and chooses the largest value that gives the minimum misclassification rate. The reason for this choice is that one can achieve maximised stability without significant loss of separability. Computing procedure for the PDV method was described in detail in Reference 11.

Results and discussion

NIR spectra of whole blood samples from mastitic and healthy cows were almost identical with the spectrum of water. The spectra from two different classes are overlapped severely throughout the whole spectral region.[11] PDV, PCA, DPLS, SIMCA, and FLDA were used to discriminate the NIR spectra of blood samples from two different classes. PCA and DPLS could not discriminate the two kinds of data. SIMCA yielded much better result, but still the discrimination performance was not sufficient. The best discrimination with SIMCA was achieved for the NIR data when the dimensionalities were both set to 20 for Class 1 and 2. The misclassification rate of SIMCA was 5 and 10, respectively, for both the training and prediction sets. It was noted in the plot of FLDA that the training objects from two classes were overfitted and positioned in two clearly separated line segments and the scores of prediction objects were scattered with a large variability so that the prediction objects from different classes yielded scores that overlapped severely.

The PDV method clearly discriminated the NIR spectra of blood samples from mastitic and healthy cows. A clear separation between the scores from different classes was obtained for the training set, but the scores of one prediction object were far-away from the training objects. This indicated that this object was an outliner in the sample set from healthy cows. One could also observe that the scores of the prediction objects were distributed in the neighbourhood of those of the training set, suggesting that the model had a desirable stability.

Conclusion

It was found that the PDV method is quite powerful for the classification of multicollinear data such as NIR spectra. The results with the NIR data of cow-blood samples showed that the PDV method yields superior performance compared to SIMCA, PCA, DPLS and FLDA. This method has opened up a new possibility in NIR spectroscopy-based classification.

References

1. G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York, USA (1992).
2. W.J. Krzanowski, *J. Chemom.* **9,** 509 (1995).
3. L. Stahle and S. Wold, *J. Chemom.* **1,** 185 (1987).
4. M. Ortiz, L. Sarabia, C. Symington, F. Santamaria and M. Iniguez, *Analyst* **121,** 1009 (1996).
5. B.K. Alsberg, D.B. Kell and R. Goodacre, *Anal. Chem.* **70,** 4126 (1998).
6. S. Wold, *Pattern Recogn.* **8,** 127 (1976).
7. O.V. Kvalheim, K. Oygard and O.G. Nielsen, *Anal. Chim. Acta* **150,** 145 (1987).
8. P.J. Gemperline, L.D. Webber and F.O. Cox, *Anal. Chem.* **1989,** 138 (1989).
9. I.E. Frank and J.H. Friedman, *J. Chemom.* **3,** 463 (1989).
10. U.G. Indahl, N.S. Sahni, B. Kirkhus and T. Næs, *Chemom. Intell. Lab. Syst.* **49,** 19 (1999).
11. J.-H. Jiang, R. Tsenkova, Y. Wu, R.-Q. Yu. and Y. Ozaki, *application Spectrosc.* **56,** 488 (2002).
12. R. Tsenkova, S. Atanassova, K. Toyoda, Y. Ozaki, K. Itoh and T. Fearn, *J. Dairy Sci.* **82,** 2344 (1999).