Meat speciation using an hierarchical approach and logistic regression

Thorsteinn Arnalds,[a] Tom Fearn[a] and Gerard Downey[b]

[a]*Department of Statistical Science, University College London, London WC1E 6BT, UK*

[b]*Teagasc, The National Food Centre, Dunsinea, Castleknock, Dublin 15, Republic of Ireland*

Introduction

Speciation of fresh, comminuted meat is an important authenticity issue.[1,2] A number of approaches to this problem have been reported, including some based on chemometric analysis of mid- and near infrared spectroscopic data.[3,4] Levels of success achieved in these feasibility studies, which involved discrimination between selected meat species (chicken, turkey, pork, beef and lamb) have been encouraging, although not sufficiently accurate to warrant their immediate use by regulatory agencies or the food industry. Given the obvious practical advantages of spectroscopic techniques, there is a strong interest in the evaluation of alternative chemometric classification strategies to address this issue.

Techniques previously investigated have included factorial discriminant analysis (FDA), *k*-nearest neighbours analysis (K-NN), partial least squares regression (PLSI & PLSII) and soft independent modelling of class analogy (SIMCA). Each of these has its own advantages and disadvantages. In the case of FDA, K-NN and PLSII, the discrimination takes place in a multivariate space defined by all of the classes to be classified. This allows a one-step model development but may present difficulties in distinguishing between all of the different sample types effectively. Additionally, when a new type of, for example, meat needs to be added to the model, it (the model) must be developed *de novo* all over again. For SIMCA, each class of material needs to be modelled separately; addition of a new class is straightforward and quick. PLSI regression requires all of the sample types to be present during model development and necessitates a separate model for each material class.

The work reported in this paper describes two other approaches to general discrimination problems using a dataset previously described.[4] The first approach constructs the classification problem as a hierarchy of binary decisions, the correct solution to each leading to the correct identification of an unknown. The second feature lies in the construction of the decision-making rule applied at each step—this uses a technique called logistic regression. Essentially, logistic regression establishes membership of one or more groups on the basis of a probability function[5] rather than the value of a predicted dummy variable or distance function. It may be applied to two (binary) or more than two (polychotomous) groups[6]—this report considers binary regression only. The hierarchical and logistic regression approaches are applied after factorisation by PCA, PLSI and PLSII.

Materials and methods

Meat samples

Two hundred and thirty (230) homogenised meat samples were utilised in this study. They comprised 55 chicken, 54 turkey, 55 pork, 32 beef and 34 lamb. Chicken and turkey were purchased as

Near Infrared Spectroscopy: Proceedings of the 10th International Conference

breast meat, pork as loin chops, beef as round steak and lamb as side loin chops; all were stored overnight at +4°C following purchase and prior to preparation and spectral collection. Individual samples were cut into cubes of manageable size and homogenised (Robot Coupe SA, Vincennes, France).

Spectral collection

Combined visible and near infrared spectra were collected in reflectance mode using an NIRSystems 6500 instrument (NIRSystems Inc., Maryland, USA) over the wavelength range 400–2500 nm at 2 nm intervals. Spectrophotometer control and spectral file management were performed using NIRS3 software (version 3.10; ISI International, Port Matilda, USA).

Chemometric procedures

The development of FDA, K-NN, SIMCA and PLS regression models for this dataset have been described previously.[4] Logistic regression models were developed from sample scores obtained by principal component analysis or partial least squares. These scores were calculated in MatLab using the PLS_Toolbox 2.0[7] while the logistic regressions were performed in Splus. The sample set was divided (on the basis of alternate samples) into separate calibration development and prediction sets.

Results and discussion

Figure 1 shows the structure of the decision making process. An unknown sample is initially classified as either red meat or white; thereafter, appropriate binary decisions allow its eventual identification. This approach utilises the natural structure of the dataset.

Using PCA factorisation, the results obtained for each of the decision steps and the overall classification success is summarised in Table 1. Initial segregation into red or white meat classes is done on the basis of scores on components 2 and 3; this is achieved with 100% success. Complete discrimination between beef and lamb meats was similarly achieved using sample scores from PCs 5 and 11. In the case of white meat, the first decision is between pork and poultry meats. In this case, a logistic regression model (fitted using stepwise forward regression) comprising four principal components (4,9,15 and 18) proved optimum. In the calibration sample set, one poultry sample was mis-classified as pork while, in the prediction set, two poultry samples were mis-classified as pork and three pork samples as poultry. Discrimination between chicken and turkey samples was problematic as discovered previously.[3,4] A logistic regression model involving three components mis-classified 13 samples overall—three in calibration and ten in prediction. These latter were all chicken.

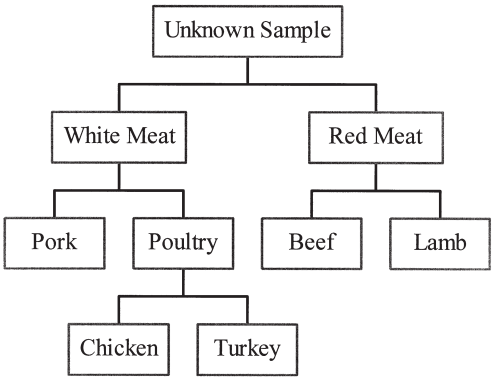

Figure 1. Schematic representation of the hierarchical classification approach.

In the case of PLS factorisation, results are summarised in Tables 2 and 3. As for PCA factorisation, the first two decisions were made most effectively with a linear cut-off rather than a logistic regression model. Subsequent decisions were best made with a logistic regression approach. It can be seen from Table 2, that PLSI factorisation produced perfect classification in the calibration sample sets but that predictive performance with regard to pork vs poultry and especially turkey vs chicken was disappointing. This latter reduced the overall correct classification rate in prediction for the five groups to 78.3%, the lowest of the three factorisation methods. PLSII factorisation produced results which

Table 1. Summary of classification results using PCA factorisation.

			% Correct Classification	
Decision	PCs	Method	Calibration Set	Prediction Set
Red vs white	2, 3	linear cut-off	100	100
Lamb vs beef	5, 11	linear cut-off	100	100
Pork vs poultry	4, 9, 15, 18	logistic	98.8	93.9
Total within 4 groups			*99.1*	*95.7*
Turkey vs chicken	13, 6, 9	logistic	94.4	81.8
Total within 5 groups			*96.5*	*87.0*

Table 2. Summary of classification results using PLSI factorisation.

			% Correct Classification	
Decision	PCs	Method	Calibration Set	Prediction Set
Red vs white	2	point cut-off	100	100
Lamb vs beef	2,7	point cut-off	100	97.0
Pork vs poultry	2, 1 ,4 ,9	logistic	100	90.2
Total within 4 groups			*100*	*92.2*
Turkey vs chicken	3, 4, 6, 15, 16	logistic	100	70.9
Total within 5 groups			*100*	*78.3*

Table 3. Summary of classification results using PLSII factorisation.

			% Correct Classification	
Decision	PCs	Method	Calibration Set	Prediction Set
Red vs white	2	point cut-off	100	100
Lamb vs beef	4	point cut-off	100	87.9
Pork vs poultry	5, 13, 11 ,3	logistic	98.8	93.9
Total within 4 groups			*99.1*	*92.2*
Turkey vs chicken	6, 8, 10, 15	logistic	96.3	85.5
Total within 5 groups			*97.4*	*85.2*

were better than PLSI. Overall, the best performance obtained in this study was using the PCA factorisation approach.

A summary of the results obtained here and those previously reported[3] is shown in Table 4. The PCA factorisation methods which performed best in this work compared favourably with factorial discriminant analysis. The PLSI technique produced the poorest results. In all cases, the models had

Table 4. Summary of % correct classification results using several chemometric techniques.

	Four groups			Five groups		
Technique	Cal	Val	All	Cal	Val	All
FDA[3]	100	95.7	97.8	91.3	86.1	88.7
K-NN[3]	91.2	86.1	88.7	87.0	77.4	82.2
PCA factorisation	99.1	95.7	97.4	96.5	87.0	91.7
PLSI factorisation	100	92.2	96.1	100	78.3	89.1
PLSII factorisation	99.1	92.2	95.7	97.4	85.2	91.3

greatest difficulty distinguishing between chicken and turkey. One of the concerns arising from this work is the high probabilities often associated with mis-classified samples i.e. the estimated probability of a sample belonging to an incorrect class is often close to 1.0 and *vice versa*. This is thought to arise from the intrinsic structure of the dataset being modelled and may indicate a limitation in the utility of logistic regression in this type of application.

Conclusions

The approach to classification described in this report has produced models of comparable accuracy to the best previously published. With regard to the hierarchical approach, it uses the inherent structure in the data and makes the decision-making process transparent. Its success in the classification of red *vs* white and poultry *vs* pork meats is striking. The models produced by logistic regression are not developed using rigorous statistical procedures but on the basis of results obtained and this is a potential weakness. No attempt has been made to optimise the classification results through, for example, data pre-treatment or variable selection. The hierarchical approach has advantages for such optimisations since different sets of variables or data treatments may easily be used for each classification step.

References

1. I.D. Lumley, in *Food Authentication*, Ed by P.R. Ashurst and M.J. Denis. Blackie Academic & Professional, London, UK, pp 108–139 (1996).
2. K.D. Hargin, *Meat Sci.* **43(S),** 277 (1996).
3. H. Rannou and G. Downey, *Analytical Communications* **34,** 401(1997).
4. J. McElhinney, G. Downey and T. Fearn, in *Near Infrared Spectroscopy: Proceedings of the 9th International Conference,* Ed by A.M.C. Davies and R. Giangiacomo. NIR Publications, Chichester, UK, pp. 511–515 (2000).
5. A.J. Dobson, *An Introduction to Generalised Linear Models.* Chapman & Hall/CRC, Boca Raton, USA (1990).
6. C.B. Begg and R. Gray, *Biometrika* **71(1),** 11 (1984).
7. B.M. Wise and N.B. Gallagher, *PLS_Toolbox 2.0* for use with MatLab. Eigenvector Research Inc. (1998).