PE4.108 Statistical classification techniques for distinguishing sheep category based on fatty acid profiles
396.00

Peter Watkins (1,2) Peter.Watkins@dpi.vic.gov.au, David Clifford(3), David Allen (2), Gavin Rose (2), Robyn Warner (2), Frank Dunshea (4) David Pethick (1)
(1)Murdoch University
(2)Department of Primary Industries
(3)CSIRO Mathematical & Information Science
(4)University of Melbourne

***Abstract*—Dentition is used as a proxy for age for defining meat quality in Australian sheep meat. This approach, though, might be inaccurate. Thus, the availability of an objective method for determining sheep age (and thus category) would potentially remove any inaccuracies. Statistical classification algorithms have been successfully used in bioinformatics. We evaluated this approach for determining sheep age focusing on the performance of three statistical algorithms (support vector machines, recursive partitioning and random forests). The algorithms were applied to the measured fatty acid profiles of fat samples from 533 carcasses; 254 lambs (< 1 year old), 131 hogget (about 1-2 years old) and 148 mutton (> 2 years old) samples. Three data pre-treatments (range transformation, column mean centering and range transformation with mean centering) were also examined to determine their impact on the performance of the algorithms. The random forests algorithm, when applied to mean-centered data, gave 100 % predictive accuracy when classifying sheep category. This approach could be used for the development of an objective test for determining sheep category.**

P. Watkins is with the Victorian Department of Primary Industries, Werribee, 3030, Australia, presently on study leave from CSIRO Food Science Australia (corresponding author: +613-97428621; fax:+613-97420400; e-mail: Peter.Watkins@dpi.vic.gov.au)

D. Clifford is with CSIRO Mathematical and Information Sciences, North Ryde, 1670 NSW Australia (e-mail: David.Clifford@csiro.au)

D. Allen is with the Victorian Department of Primary Industries, Werribee, 3030, Australia (e-mail: David.Allen@dpi.vic.gov.au)

G. Rose is with the Victorian Department of Primary Industries, Werribee, 3030, Australia (e-mail: Gavin.Rose@dpi.vic.gov.au)

R. Warner is with the Victorian Department of Primary Industries, Werribee, 3030, Australia (e-mail: Robyn.Warner@dpi.vic.gov.au)

F. Dunshea is with the Melbourne School of Land and Environment, University of Melbourne, Parkville, 3051 Victoria Australia (e-mail: frdunshea@unimelb.edu.au).

D. Pethick is with the School of Veterinary and Biomedical Sciences, Murdoch University, Murdoch, 6015 WA Australia, (e-mail: D.Pethick@murdoch.edu.au).

***Index Terms*— random forest, sheep age, statistical classification, fatty acids.**

I. Introduction

The Australian sheep meat industry uses dentition as a proxy for age for defining quality with the categories of lamb (no erupted permanent incisors), hogget (2 erupted incisors) or mutton (greater than 2 erupted incisors) [1]. Recent work though suggests that this practice may be inaccurate since, for a flock of research sheep, it was found that the age of eruption for permanent incisors ranged from 369 to 483 days, with differences also evident across breeds [2]. Recently, an Australian Senate inquiry has reported that there is some concern that substitution of lamb with older animals may be occurring within the industry [3]. Given that Australians spend $AU 2 billion annually to purchase lamb (regarded by consumers as a premium meat product), it would be useful if an objective measuring tool for determining sheep age were available to remove any concern that meat substitution might be occurring. Branched chain fatty acids (BCFAs) are the main compounds responsible for mutton flavour/odour found in the cooked meat of older sheep and can be used to differentiate lamb from mutton but only if the pre-slaughter nutritional history is known [4]. Measuring the BCFA content of sheep fat is an objective way for determining age (and thus category) yet details on pre-slaughter nutrition for an animal may not always be available. Thus, it would be useful if there was an alternative approach for classifying sheep age which did not need any additional information. Recently, developments in bioinformatics have seen the use of statistical algorithms as classifiers that can be used to distinguish between two exclusive states (e.g. differences between 'normal' *vs.* 'abnormal' cells). This approach suggested itself as one which might be suitable for distinguishing sheep category, using the measured fatty acid profiles that we have obtained elsewhere [4]. Thus, we investigated whether statistical algorithms could be used for classifying sheep category; more specifically, our aim was to determine whether three

specific algorithms (support vector machines, recursive partitioning and random forests) were suitable for distinguishing between lamb, hogget and mutton.

II. MATERIALS AND METHODS

A. Sample preparation and analysis

Full details of sample preparation and analysis are given in an accompanying paper [4]. Briefly, fat samples from 254 lamb, 131 hogget and 148 mutton fat samples were taken from animals of different age, breed and sex in three states of Australia; Victoria, New South Wales and Western Australia. Molten fat (1g) was heated in a Unitrex co-distillation unit and the released fatty acids (FA) were collected and derivatised as trimethylsilyl (TMS) esters. The FA-TMS esters were separated using a Varian 3400 gas chromatograph and detected using a Varian Saturn 2000 ion trap mass spectrometer operating in full scan mode. The measured profile, as total abundance against retention time, formed a total ion chromatogram (TIC).

B. Statistical classification

Each TIC was exported from the Varian Star Workstation software for use with R [5]. Over the period of data acquisition, the retention times of some peaks in the TICs had changed which meant that the chromatograms had to be aligned. This was done using variable penalty dynamic time warping [6]. After background removal using asymmetric least squares [7], the variable penalty and the master signal to which the TICs would be aligned were first defined, and each TIC was aligned to the master signal. The TICs were stacked as a matrix that contained the combined data set as 2749 columns (time points) and 533 rows. Three data pre-treatments were applied to the matrix. These were a) range transformation of the TIC between 0 and 100, b) column mean centering of the untreated data and (c) column mean centering of the range transformed data. Range transformation of the TIC was done using

$$x_i^* = (x_i - x_{\min})/(x_{\max} - x_{\min}) \text{ X } 100$$

where x_i^* is the scaled data for each row i, x_i is the measured TIC response, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the TIC. Column mean centering of the data was performed using

$$x_{ik}^* = x_{ik} - \bar{x}_k$$

where x_{ik}^* is the treated data entry for row i and column k, and $\bar{x}_k$ is the column mean.

Three different statistical algorithms were applied to the three transformed data sets as well as the original data set. The algorithms were support vector machines (SVM), recursive partitioning (RP), and random forests (RF). The efficacy of each algorithm to classify sheep category was tested using 10-fold cross-validation. This meant that the data set was partitioned into approximately 10 equal-sized parts and, for the kth part ($1 \leq k \leq 10$), the model was fit to the other k-1 parts of the data [7]. The kth part of the dataset was the test set while the remaining data was the training set. The error of the model for predicting the kth part of the data was then calculated. This was done for all k = 1,…, 10 with the final result taken as the average of all 10 estimates of prediction error.

III. RESULTS AND DISCUSSION

Figure 1 shows all the TICs for the different sheep categories stacked beside each other as an image, before alignment (upper panel) and after variable penalty dynamic time warping (DTW). In the upper panel, the measured profiles for lamb (in order of acquisition date) are shown in the panel's lower half, while hogget and mutton can be seen as two distinct groups in the top half of the panel. For each category, it was evident that shifts in the retention times had occurred over the period of data acquisition. Such shifts usually result from changes in the instrumental operating conditions [9]. Variable penalty DTW [6] was used for peak alignment of the data entire set. Almost all of the peaks in the data set were aligned except for the last eluting compound which was identified as the TMS ester of cholesterol. Cholesterol, a natural component of ovine fat [10], was not expected to be significantly involved with differentiating sheep category.

The performance of three classification algorithms (support vector machines, recursive partitioning and random Forests) on the four datasets was evaluated (Table 1). Of the three algorithms, random forests was the best performer with 100 % accuracy with the mean centered data compared to ~ 85 % accuracy with the original and range transformed data with recursive partitioning as the least accurate algorithm in classifying and support vector machines as an intermediate.

Column mean centering and range transformation are two scaling methods that can be used for pre-treatment of a data set [11]. Often the differences between measured features in a data set can be in orders of magnitude, resulting in a higher influence of some variables compared to the remainder. For this work, the main components of the TIC were hexadecanoic (palmitic, $C_{16:0}$), octadecenoic (oleic, $C_{18:1}$) and octadecanoic (stearic, $C_{18:0}$) FA-TMS esters [4]. This result was not unexpected since these FAs are the major ones in ovine fat, ranging from ~ 20 to 30 g per 100g of the total fatty acid content [12]. Given the abundance of these FAs in the samples, mean centering and range transformation were used to pre-treat data as it was anticipated that these techniques would reduce the influence of these components in any subsequent data analysis. Of the two, only column mean centering was the most effective pre-treatment step with very little error found for each algorithm in predicting sheep category. Mean centering is regarded as a standard pre-treatment technique in chemometric analysis [13].

In this study, we have used the complete TIC as a "chemical fingerprint" of sheep category for subsequent multivariate analyses. This approach implies that the nature and relative amount of each compound are distinctive features of the TIC and the associated sample. While feature selection could be applied to the TICs to significantly reduce the size of the dataset [8], this represents a "reductionist" approach which may only identify a few peaks to the exclusion of other possible candidate compounds that could be suitable for classification. The use of the whole TIC represents a "holistic" approach, one that has been advocated in modern systems biology, and gaining widespread application in that discipline [14]. Using the complete chromatogram does introduce an additional computational penalty to the data analysis but, given the high processing performance associated with modern PCs, this does not represent too high an impost on the analysis.

The random forests algorithm was the best classifier for predicting sheep type and could be useful as an objective means for determining sheep category. The current practice of sheep classification relies on dentition; i.e. counting the number of an animal's teeth prior to slaughter. Currently, there are no objective methods for determining sheep category that can be used at either the processing stage or at any subsequent point in the supply chain. If such an approach could be developed using, say, random forests then there is a possibility that meat substitution could be detected, and so help to minimise this practice in industry.

IV. CONCLUSION

The fatty acid profiles of 533 sheep fat samples were measured and three statistical classifiers were applied to the measured profiles. Three data pre-treatment techniques were also applied to the data. The highest accuracy (100%) for predicting sheep age was found with the random forests classifier and column mean centering for pre-treating the data. The statistical approach could be used as an objective test for determining sheep category. The availability of such a test has implications for the sheep meat industry as it could be used for detecting meat substitution and assist in reducing this practice within the industry.

ACKNOWLEDGEMENT

Funding by the Co-operative Research Centre for Sheep Industry Innovation is gratefully acknowledged.

REFERENCES

[1]. Pethick, D.W., Hopkins, D.L., D'Souza, D.N., Thompson, J.M., & Walker, P.J. (2005) Effects of animal age on the eating quality of sheep meat. Australian Journal of Experimental Agriculture 45, 491-498.

[2]. Hopkins, D.L., Stanley, D.F., Martin, L.C., & Gilmour AR (2007) Genotype and age effects on sheep meat production. 1. Production and growth. Australian Journal of Experimental Agriculture 47, 1119-1127.

[3]. Senate Standing Committee on Rural and Regional Affairs and Transport. (2008). Meat marketing – interim report. Canberra: Parliament of Australia

[4]. Watkins, P.J, Rose, G., Allen, D., Salvatore, L., Warner, R.D., Dunshea, F.R., & Pethick, D.W. (2009). Chemical basis for discriminating lamb from mutton. In Proceedings of the 55th International Congress of Meat Science and Technology, 16-21 August, Copenhagen, Denmark.

[5]. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[6]. Clifford, D., Stone, G., Montoliu, I., Rezzi S., Martin, F.P., Guy, P., Bruce, S., & Kochhar, S. (2009). Alignment using variable penalty dynamic time warping. Analytical Chemistry 81, 1000-1007.

[7]. Eilers, P.H.C. (2004). Parametric time warping. Analytical Chemistry, 76, 404-411.

[8]. Hastie, T., Tibshiranii, R., & Friedman, J. (2009). 'The Elements of Statistical Learning: Data Mining, Inference, and Prediction.' (2nd edn) New York: Springer.

[9]. Chae, M., Shmookler Reis. R.J., & Thaden, J. (2008). An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas chromatography-mass spectrometry peaks. BMC Bioinformatics 9, S15.

[10]. Nelson, G.J. (1967). Composition of neutral lipids from erythrocytes of common mammals. Journal of Lipid Research, 8, 374-379.

[11]. Berrueta, L.A., Alonso-Salces, R.A., & Héberger, K. (2007). Supervised pattern recognition in food analysis. Journal of Chromatography A, 1158, 196-214.

[12]. Wood, J.D., Enser, M., Fisher, A.V., Nute, G.R., Sheard, P.R., Richardson, R.I., Hughes, S.I., & Whittington, F.M. (2008). Fat deposition, fatty acid composition and meat quality: a review. Meat Science 78, 343-358.

[13]. Flåten, G.R., & Walmsley, A.D. (2003). Using design of experiments to select optimum calibration model parameters. Analyst 128, 935-943.

[14]. Kell, D.B. (2004). Metabolomics and systems biology: making sense of the soup. Current Opinion in Microbiology 7, 296-307.

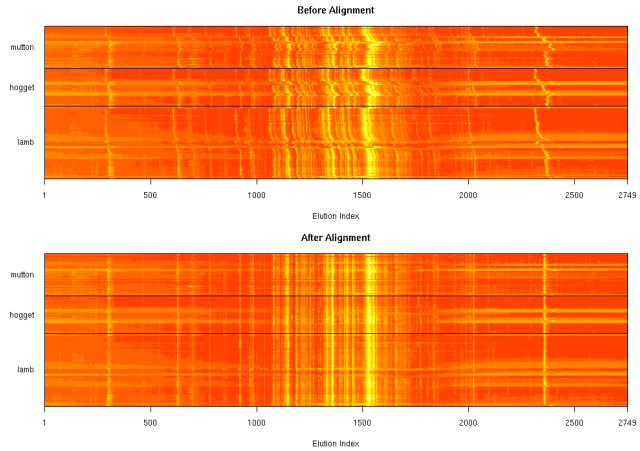

Figure 1. Coloured scale representation of aligning total ion chromatograms, based on alignment using variable penalty dynamic type warning. Intensity is proportional to the logarithm of measured abundance. Top panel: before alignment. Bottom panel: after alignment.

Table 1. Performance of classification algorithms for discriminating lamb, hogget and mutton (as % accuracy)

	Data pre-treatment			
Classifier	A	B	C	D
SVM	70	78	99	99
RP	66	69	95	95
RF	84	86	100	100

Classifiers were: SVM = support vector machines, RP = recursive partitioning, RF = random forest. Data pretreatment: A = original data B = range transformed data C = column mean centering of original data D = column mean centering of range transformed data

.