Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 89 (2007) 102–115
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies

Deirdre Toher [a,b,*], Gerard Downey [a], Thomas Brendan Murphy [b]

[a] *Ashtown Food Research Centre, Teagasc, Dublin 15, Ireland*
[b] *Department of Statistics, School of Computer Science and Statistics, Trinity College, Dublin, Dublin 2, Ireland*

Received 10 July 2006; received in revised form 8 May 2007; accepted 12 June 2007
Available online 22 June 2007

Abstract

Classification methods can be used to classify samples of unknown type into known types. Many classification methods have been proposed in the chemometrics, statistical and computer science literature.

Model-based classification methods have been developed from a statistical modelling viewpoint. This approach allows for uncertainty in the classification procedure to be quantified using probabilities. Linear discriminant analysis and quadratic discriminant analysis are particular model-based classification methods.

Partial least squares discriminant analysis is commonly used in food authentication studies based on spectroscopic data. This method uses partial least squares regression with a binary outcome variable for two-group classification problems.

In this paper, model-based classification is compared to partial least squares discriminant analysis for its ability to correctly classify pure and adulterated honey samples when the honey has been extended by three different adulterants. Two model selection criteria are examined: the Bayesian Information Criterion and 5-fold cross validation. The methods are compared using the classification performance and the interpretability of the results.

In addition, since the percentage of adulterated samples in any given sample set is unlikely to be known in a real-life setting, the ability of updating procedures within model-based clustering to accurately predict the adulterated samples, even when the proportion of pure to adulterated samples in the training data is grossly unrepresentative of the true situation, is studied in detail.

The performance of both model-based and partial least squares discriminant analysis is found to be robust to the composition of the training data and to model selection method. The Bayesian Information Criterion is shown to be more robust than 5-fold cross validation as a model selection method, especially when the training data set is very small and unrepresentative of the entire data set.

Keywords: Food authenticity; NIR spectroscopy; Classification; Model-based classification; Partial least; squares regression

1. Introduction

The main aim of food authenticity studies is to detect when foods are not what they claim to be and thereby prevent economic fraud or possible damage to health. Foods that are susceptible to such fraud are those which are expensive and subject to the vagaries of weather during growth or harvesting e.g. coffee, various fruits, herbs and spices. Food fraud can generate significant amounts of money (e.g. several million US dollars) for unscrupulous traders so the risk of adulteration is real. Honey is defined by the EU [1] as "the natural, sweet product produced by Apis mellifera bees from the nectar of plants or from secretions of living plants, which bees collect, transform by combining with specific substances of their own, deposit, dehydrate, store and leave in honeycombs to ripen and mature". As it is a relatively expensive product to produce and extremely variable in nature, honey is prone to adulteration for economic gain. Instances of honey adulteration have been recorded since Roman times when concentrated grape juice was sometimes added, although nowadays industrial syrups are more likely to be used as honey extenders. False claims may

* Corresponding author. Department of Statistics, School of Computer Science and Statistics, Trinity College, Dublin, Dublin 2, Ireland. Tel.: +353 1 8059500; fax: +353 1 8059550.
E-mail address: toherd@tcd.ie (D. Toher).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.06.005

also be made in relation to the geographic origin of the honey but this study concentrated on attempting to classify samples as either pure or adulterated. In this study, artisanal honeys were adulterated in the laboratory using three adulterants – fructose: glucose mixtures, fully-inverted beet syrup and high fructose corn syrup – in various ratios and weight percentages.

Model-based classification [2] is a classification method based on the Gaussian mixture model with parsimonious covariance structure. This method models data within groups using a Gaussian distribution and the abundance of each group has some fixed probability. This classification method has been shown to give excellent classification performance in a wide range of applications [3]. A recent extension of model-based classification that uses data with unknown group membership in the model-fitting procedure has been developed [6]. A detailed review of model-based classification and its extensions is given in Section 4.1.

Partial least squares regression is a method that seeks to optimise both the variance explained and correlation with the response variable [9]. In a previous study [10] it was found to outperform other chemometric methods commonly used in the study of near-infrared transflectance spectra. It has the advantage in that it can utilise highly-correlated variables for classification purposes.

Both model-based classification and partial least squares discriminant analysis requires training on data with known group or class labels. The collection of training data in food authenticity can be very expensive and time-consuming so methods that require few training data observations are particularly useful.

We show that both methods give excellent classification performance, even when few training data values are available. We also find that model-based classification is robust in situations where the training and test data are quite different.

Model-based classification, along with updating procedures were previously found [6] to outperform partial least squares regression on food science spectroscopy data. However, we discover in this comparison paper that such updating procedures are not always beneficial. There are recognised connections between partial least squares and canonical discriminant analysis as described by [11]; however the approach taken by model-based discriminant analysis provides a more flexible approach to modelling the shape of the groups.

2. Materials and methods

2.1. Honey samples

Honey samples (157 samples) were obtained directly from beekeepers through-out the island of Ireland. Samples were from the years 2000 and 2001; they were stored unrefrigerated from time of production and were not filtered after receipt in the laboratory. Prior to spectral collection, honeys were incubated at 40°C overnight to dissolve any crystalline material, manually stirred to ensure homogeneity and adjusted to a standard solids content (70° Brix) to avoid spectral complications from naturally-occurring variations in sugar concentration.

Collecting and extending the honey and recording the spectra was done at time points several months apart; the first study involved extending some of the authentic samples of honey with fructose:glucose mixtures, the second study involved extending some of the remaining authentic samples with fully-inverted beet syrup and high fructose corn syrup. All adulterant solutions were also produced at 70° Brix. Brix standardisation of honeys and adulterant solutions meant that any adulteration detected would not be simply on the basis of gross added solids.

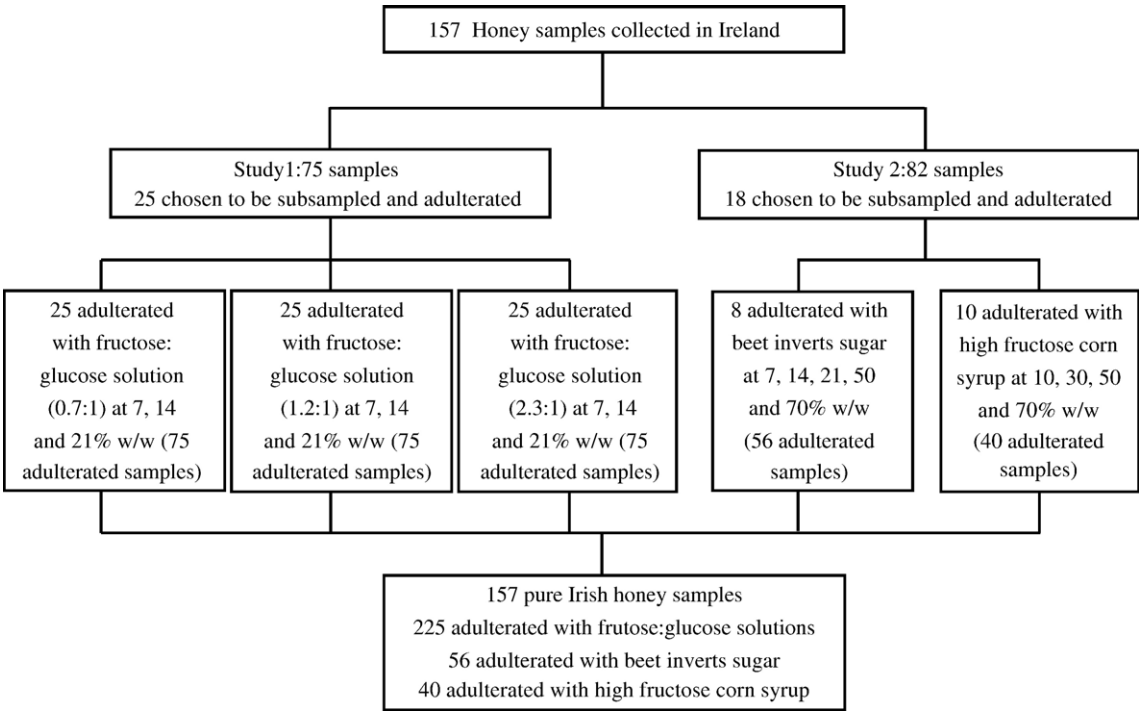

Fig. 1. Flow diagram of adulteration process.