Evolutionary Approaches to Variable Selection

Marcos Gestal
University of A Coruña, Spain

José Manuel Andrade
University of A Coruña, Spain

INTRODUCTION

The importance of juice beverages in daily food habits makes juice authentication an important issue, for example, to avoid fraudulent practices.

A successful classification model should address two important cornerstones of the quality control of juice-based beverages: to monitor the amount of juice and to monitor the amount (and nature) of other substances added to the beverages. Particularly, sugar addition is a common and simple adulteration, though difficult to characterize. Other adulteration methods, either alone or combined, include addition of water, pulp wash, cheaper juices, colorants, and other undeclared additives (intended to mimic the compositional profiles of pure juices) (Saavedra, García, & Barbas, 2000).

VARIABLE SELECTON BY MEANS OF EVOLUTIONARY TECHNIQUES

This chapter presents several approaches to address the variable selection problem. All of them are based on evolutionary techniques. They can be divided into two groups. First group of techniques are based on different codifications of a traditional Genetic Algorithm (GA) population and different specifications for the evaluation function. Second group shows a modification in the traditional Genetic Algorithm to improve the generalization capability by adding a new population and an approach based on the evolution of subspecies into the genetic population.

BACKGROUND

A range of analytical techniques have been used to deal with authentication problems. These include high performance liquid chromatography (Yuan & Chen, 1999), gas chromatography (Stöber, Martin & Peppard, 1998) and isotopic methods (Jamin, González, Remaud, Naulet & Martin, 1997). Unfortunately, they are expensive and slow.

Infrared Spectrometry (IR) (Rodriguez-Saona, Fry, McLaughlin, & Calvey, 2001) is a fast and convenient technique to perform screening studies in order to assess the quantity of pure juice in commercial beverages. The interest lies in developing, from the spectroscopy data, classification methods that might enable the determination of the amount of natural juice contained in a sample.

However, the information gathered from the IR analyses has some fuzzy characteristics (random noise, unclear chemical assignment, etc.), so analytical chemists tend to use techniques like Artificial Neural Networks (ANN) (Haykin, 1999) or develop ad-hoc classification models. Previous studies (Gestal, Gómez-Carracedo, Andrade, Dorado, Fernández, Prada, & Pazos, 2005) showed that ANN classified apple juice beverages according to the concentration of natural juice they contained and that ANN had advantages over classical statistical methods, such as robust models and easy application of the methodology on R&D laboratories. Disappointingly, the large number of variables derived from IR spectrometry makes ANNs time-consuming during training and, most important, makes it very difficult to establish relationships between these variables and the analytical knowledge.

Several approaches were used to reduce the number of variables to a small subset, which should retain the classification capabilities of the overall dataset. Hence, the ANN training process and the interpretation of the results would be highly improved.

Furthermore, previous variable selection would yield other advantages: cost reduction (if the classification model requires a reduced set of data, the time needed to obtain them will be shorter; increased efficiency (if the system processes less information, less

time for processing it will be required); understanding improvement (if two models resolve the same task, but one of them uses less information this would be more thoroughly interpreted. Therefore, the simpler the model, the easier the knowledge extraction and the easier the understanding, the easier the validation).

In addition, it was proved the analysis of IR data involved a highly multimodal problem, as many combinations of variables each (obtained using a different method) led to similar results when the samples were classified.

GENETIC ALGORITHMS

A GA (Holland, 1975)(Goldberg, 1989) is a recurrent and stochastic process that operates with a group of potential solutions to a problem, known as genetic population, based on one of the Darwin's principles: the survival of the best individuals (Darwin, 1859).

Briefly a GA works as follows. Initially, a population of solutions is generated randomly and the solutions evolve continuously after consecutive stages of crossovers and mutations. Every individual at the population has an associated value that quantifies associated its usefulness (adjustment or fitness), in accordance to its adequacy to solve the problem. This value has to be obtained for each potential solution and constitutes the quantitative information the evolutionary algorithm will use to guide the search. The process will continue until a predetermined stopping criterion is reached. This might be a particular threshold error for the solution or a certain number of generations (populations).

Therefore, different basic steps will be required to implement a GA: codification of the problem, which results in a population structure, initialisation of the first population, defining a fitness function to evaluate how good is each individual to solve the problem and, finally, a cyclic procedure of reproductions and replacements (Michalewicz, 1999; Goldberg, 2002).

DATA DESCRIPTION

In the present practical application, the spectral range measured by IR spectrometry (wavenumbers from 1250 cm-1 to 900 cm-1) provided 176 absorbances (which measured light absorption)(Gómez-Carracedo, Gestal, Dorado & Andrade, 2007).

The main goal of the application consisted on the prediction of the amount of pure juice on a sample using absorbance values returned for the IR measurements. But the amount of data obtained for a sample by IR spectrometry is huge, so the direct application of mathematical and/or computational methods (although possible) requires a lot of time. Accordingly, it is important to establish whether all raw data provided relevant information for sample differentiation. Hence, the problem was an appropriate case for the use of variable selection techniques.

Previous to variable selection construction of data sets for both model development and validation was required. Thus, samples with different amounts of pure apple juice were prepared at the laboratory. Besides, 23 apple juice-based beverages sold in Spain were analysed (the declared amount of juice printed out on

Table 1. Low and high concentrations dataset

Juice Concentration	2%	4%	6%	8%	10%	16%	20%	Total
Training	19	17	16	22	21	20	19	134
Validation	1	1	13	6	6	6	6	39
Commercial	0	0	0	0	1	1	0	2

Juice Concentration	20%	25%	50%	70%	100%	Total
Training	20	19	16	14	7	86
Validation	6	18	13	1	6	44
Commercial	0	2	0	0	19	21