Chapter 1.13
Several Approaches to Variable Selection by Means of Genetic Algorithms

Marcos Gestal Pose
University of A Coruña, Spain

Alberto Cancela Carollo
University of A Coruña, Spain

José Manuel Andrade Garda
University of A Coruña, Spain

Mari Paz Gómez-Carracedo
University of A Coruña, Spain

ABSTRACT

This chapter shows several approaches to determine how the most relevant subset of variables can perform a classification task. It will permit the improvement and efficiency of the classification model. A particular technique of evolutionary computation, the genetic algorithms, is applied which aim to obtain a general method of variable selection where only the fitness function will be dependent on the particular problem. The solution proposed is applied and tested on a practical case in the field of analytical chemistry to classify apple beverages.

INTRODUCTION

The main goal of any classification method is to establish either the class or the category of a given object, which is defined by some attributes or variables. Nevertheless, not all those attributes give the same quality and quantity of information when the classification is performed. Sometimes, too much information (in this chapter, this term will include both useful and redundant information) can cause problems when assigning an object to one or another class, thus deteriorating the performance of the classification.

The problem of variable selection involves choosing a subgroup of variables from an overall set of them that might carry out the finest classification. Some advantages obtained after a selection process are:

- **Cost reduction for data acquisition:** If less data are required for sample classification, the time required to obtain them would be shorter.
- **Increased efficiency of the classifier system:** Less information also requires less time for its processing.
- **Improved understanding of the classification model:** Those models that use less information to perform the same task will be more thoroughly understood. The simpler the formulation of the classifier, the easier the extraction of the knowledge and its validation.
- **Efficacy improvement:** Sometimes, too much information might deteriorate the generalization ability of the classification method.

VARIABLE SELECTION

A generic process of variable selection can be formalized by means of the following definitions:

If A is a set of n objects:

$A = \{x_i, i=1...n\}$

Each object x_i is described by a set of d variables, V, each one can be either quantitative or qualitative:

$V = \{V_j, j=1...d\}$

If C is the overall set of classes, from which discrimination is to be done, and C_k is the class to which object x_i belongs to, then any object x_i can be completely described by the set

$x_i = \{ V_{ij}, C_k\}, j=1...d$

The main goal of any variable selection procedure should be to obtain the smallest subset of variables, S ($S \subset V$), that still performs a satisfactory classification task.

Traditionally, this problem has been approached by several statistic techniques, including principal components (Gnanadesikan, 1997), cluster analysis (Jain, Murty, & Flynn, 1999), potential functions (Forina, Armanino, Learde, & Drava, 1991; Tomas & Andrade, 1999), maximum likelihood methods as SIMCA (soft independent modelling of class analogy) (Wold, Johansson, Jellum, Bjørnson, & Nesbakken, 1997), and so forth.

EVOLUTIONARY COMPUTATION

Evolutionary computation (EC) is composed of a group of techniques inspired on the biological world. As they mimic the evolutionary behaviour of the living species, they work by evolving a group of potential solutions to a given problem until the optimum solution is reached.

More formally, the term EC involves the study of the basis and application of some heuristic techniques that are based on fundamentals of natural evolution (Tomassini, 1995). This group of heuristics can be sorted into four main categories that are included in the evolutionary equation that is shown in Figure 1.

Biological Fundamentals

As it has been stated, evolutionary algorithms, in origin, tried to mimic some of the processes that take place in natural evolution. Some of the most remarkable ones are the survival of the best