

# Tabu Search for Variable Selection in Classification

**Silvia Casado Yusta**

*Universidad de Burgos, Spain*

**Joaquín Pacheco Bonrostro**

*Universidad de Burgos, Spain*

**Laura Nuñez Letamendía**

*Instituto de Empresa, Spain*

## INTRODUCTION

Variable selection plays an important role in classification. Before beginning the design of a classification method, when many variables are involved, only those variables that are really required should be selected. There can be many reasons for selecting only a subset of the variables instead of the whole set of candidate variables (Reunanen, 2003): (1) It is cheaper to measure only a reduced set of variables, (2) Prediction accuracy may be improved through the exclusion of redundant and irrelevant variables, (3) The predictor to be built is usually simpler and potentially faster when fewer input variables are used and (4) Knowing which variables are relevant can give insight into the nature of the prediction problem and allows a better understanding of the final classification model. The importance of variables selection before using classification methods is also pointed out in recent works such as Cai et al. (2007) and Rao and Lakshminarayanan (2007).

The aim in the classification problem is to classify instances that are characterized by attributes or variables. Based on a set of examples (whose class is known) a set of rules is designed and generalised to classify the set of instances with the greatest precision possible. There are several methodologies for dealing with this problem: Classic Discriminant Analysis, Logistic Regression, Neural Networks, Decision Trees, Instance-Based Learning, etc. Linear Discriminant Analysis and Logistic Regression methods search for linear functions and then use them for classification purposes. They continue to be interesting methodologies.

In this work an “ad hoc” new method for variable selection in classification, specifically in discriminant analysis and logistic regression, is analysed. This new

method is based on the metaheuristic strategy tabu search and yields better results than the classic methods (stepwise, backward and forward) used by statistical packages such as SPSS or BMDP, as it’s shown below. This method is performed for 2 classes.

## BACKGROUND

Research in variable selection started in the early 1960s (Lewis, 1962 and Sebestyen, 1962). Over the past four decades, extensive research into feature selection has been conducted. Much of the work is related to medicine and biology (e.g. Inza et al., 2002; Shy and Suganthan, 2003). The selection of the best subset of variables for building the predictor is not a trivial question, because the number of subsets to be considered grows exponentially with the number of candidate variables, which means that feature selection is a NP (Nondeterministic Polynomial)-Hard computational problem (see Cotta et al., 2004). When the size of the problem is large finding an optimum solution in practice is not feasible.

Two different methodological approaches have been developed for variable selection problems: optimal or exact techniques (enumerative techniques), which can guarantee an optimal solution, but are applicable only in small sets; and heuristic techniques, which can find good solutions (although they cannot guarantee the optimum) in a reasonable amount of time. Among the former, the Narendra and Fukunaga (1977) algorithm is one of the best known, but as Jain and Zongker (1997) pointed out, this algorithm is impractical for problems with very large feature sets. Recent references about implicit enumerative techniques of selection features adapted to regression models could be found in Gatu and

Kontoghiorghes (2005) and (2006). On the other hand, among the heuristic techniques we find works based on genetic algorithms (see Bala et al. (1996), Jourdan et al. (2001)), Oliveira et al, 2003 and Wong and Nandi, (2004)) and the recent work by Garcia et al. (2006) who present a method based on Scatter Search. As found in other optimization problems metaheuristic techniques have proved to be superior methodologies.

However, although there are many references in the literature regarding selecting variables for their use in classification, there are very few key references on the selection of variables for their use in discriminant analysis and logistic regression. For this specific purpose the *Stepwise* method (Efroymson, 1960) and all its variants, such as O’Gorman’s (2004), as well as the *Backward* and *Forward* methods, can be found in the literature. These are simple selection procedures based on statistical criteria which have been incorporated into some of the best known statistical packages such as SPSS or BMDP. As highlighted by Huberty (1994) these methods are not very efficient, and when there are many original variables the optimum is rarely achieved.

## SOLVING THE PROBLEM

### Setting Out the Problem

We can formulate the problem of selecting the subset of variables with superior classification performance as follows:  $V$  being a set of  $m$  variables, such that  $V = \{1, 2, \dots, m\}$  and  $A$  being a set of instances, (also named “training” set). For each case we also know the class it belongs to. Given a predefined value  $p \in \mathbb{N}$ ,  $p < m$ , we have to find a subset  $S \subset V$ , with a size  $p$  with the greatest classification capacity,  $f(S)$ .

To be precise, for the discriminant analysis the function  $f(S)$  is defined as a percentage of hits in  $A$  obtained through the variables of  $S$  with Fisher’s classifier (Fisher, 1936).

For logistic regression the classification capacity,  $f(s)$ , is obtained as follows: let’s consider  $S = \{1, 2, \dots, p\}$  without loss of generality; let  $x_{ij}$  be the value of variable  $j$  for case  $i$ ,  $i \in A$ ,  $j \in S$ , and  $y_i = 1$  if case  $i$  belongs first class and 0 otherwise; this classifier is obtained by calculating vector  $\mathbf{c} = (c_0, c_1, \dots, c_p)$  by maximizing the next expression

$$L(\mathbf{c}) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

where

$$p_i = \frac{1}{1 + e^{-z_i}}$$

and  $z_i = c_0 + c_1 x_{i1} + c_2 x_{i2} + \dots + c_p x_{ip}$ ;

then, once vector  $\mathbf{c}$  is obtained, every case  $i$  is classified to class 1 if  $p_i > 0.5$  and in class 2 otherwise. Function  $L$  is usually named as likelihood function.

### Solution Approach: Tabu Search

Tabu Search (TS) is a strategy proposed by Glover (1989 and 1990). A comprehensive tutorial on Tabu Search can be found in Glover and Laguna (2002).

Our Tabu Search algorithm includes a method for building an initial solution and a basic procedure for exploring the space of solutions around this initial solution. The performance of the Tabu Algorithm is outlined as follows:

#### Tabu Search Procedure

*Build an initial solution*

*Repeat*

*Execute Basic Tabu Search*

*until a stopping condition is reached*

In this work a limit of 30 minutes of computational time is used as stopping condition. Next these elements are described.

#### Initial Solution

The initial solution is built as follows: starting from the empty initial solution, a variable is added in each iteration until the solution  $S$  reaches  $p$  variables ( $|S| = p$ ). To decide which variable is added to the solution in each iteration the value of  $f$  is used.

#### Description of a Basic Algorithm

Our Tabu Search algorithm uses neighbouring moves which consist in exchanging an element that is in solu-

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/chapter/tabu-search-variable-selection-classification/11080](http://www.igi-global.com/chapter/tabu-search-variable-selection-classification/11080)

## Related Content

---

### Data Mining for Fraud Detection System

Roberto Marmo (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 411-416).  
[www.irma-international.org/chapter/data-mining-fraud-detection-system/10853](http://www.irma-international.org/chapter/data-mining-fraud-detection-system/10853)

### Knowledge Acquisition from Semantically Heterogeneous Data

Doina Caragea and Vasant Honavar (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1110-1116).  
[www.irma-international.org/chapter/knowledge-acquisition-semantically-heterogeneous-data/10960](http://www.irma-international.org/chapter/knowledge-acquisition-semantically-heterogeneous-data/10960)

### Sequential Pattern Mining

Florent Masseglia, Maguelonne Teisseire and Pascal Poncelet (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1800-1805).  
[www.irma-international.org/chapter/sequential-pattern-mining/11062](http://www.irma-international.org/chapter/sequential-pattern-mining/11062)

### Ethics of Data Mining

Jack Cook (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 783-788).  
[www.irma-international.org/chapter/ethics-data-mining/10909](http://www.irma-international.org/chapter/ethics-data-mining/10909)

### Clustering Analysis of Data with High Dimensionality

Athman Bouguettaya and Qi Yu (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 237-245).  
[www.irma-international.org/chapter/clustering-analysis-data-high-dimensionality/10827](http://www.irma-international.org/chapter/clustering-analysis-data-high-dimensionality/10827)