Available online at www.sciencedirect.com
ScienceDirect
Computers & Industrial Engineering 50 (2006) 296–311
computers & industrial engineering
www.elsevier.com/locate/dsw

An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems

Hasan Bal *, H. Hasan Örkcü, Salih Çelebioğlu

Department of Statistics, Faculty of Arts and Sciences, Gazi University, Teknikokullar, 06500, Ankara, Turkey

Received 14 March 2006; received in revised form 29 May 2006; accepted 7 June 2006
Available online 24 July 2006

Abstract

The aim of this article is to consider a new linear programming and two goal programming models for two-group classification problems. When these approaches are applied to the data of real life or of simulation, our proposed new models perform well both in separating the groups and the group–membership predictions of new objects. In discriminant analysis some linear programming models determine the attribute weights and the cut-off value in two steps, but our models determine simultaneously all of these values in one step. Moreover, the results of simulation experiments show that our proposed models outperform significantly than existing linear programming and statistical approaches in attaining higher average hit-ratios.

Keywords: Goal programming; Linear programming; Classification; Discriminant analysis

1. Introduction

Discriminant analysis has been successfully used for many fields, health applications, education planning, taxonomy problems, including engineering applications. Discriminant analysis is a technique which is interested in determining the groups of objects based on their observed scores. Especially Fisher's linear discriminant function is the most popular technique which is frequently used for the discriminant problem. As an alternative for the examination of classification problems using the statistical methods, it is recently developed a number of new efficient mathematical programming approaches. See Bajgier and Hill (1982); Erenguc and Koehler (1990); Fred and Glover (1981a, 1981b, 1986); Glover (1990); Joachimsthaler and Stam (1988); Koehler and Erenguc (1990); Lam and Moy (1997); Lam and Moy (2002); Lee and Ord (1990); Rubin (1990); Sueyoshi (1999) among others. In two-group and multigroup classification problems, Lam, Choo, and Moy (1996) and Lam and Moy (1996) developed a satisfactory

* Corresponding author.
E-mail address: hasanbal@gazi.edu.tr (H. Bal).

0360-8352/$ - see front matter
doi:10.1016/j.cie.2006.06.002

model for classification based on cluster analysis. But, in two-group classification problems, their approach minimizes the sum of deviations of all object's classification scores from the mean group classification scores.

In statistics, it is known that the median is the point minimizing the total ℓ_1-norm distance while the mean is the point that minimizes the total ℓ_2-norm distance from all points to it (Benjamin et al., 2005; Bradley, Mangasarian, & Street, 1997; Gilani & Padberg, 2002). Our suggested models and Lam et al. (1996) model, which we will call as the LCM, base on the ℓ_1-norm.

For these reasons, it is more appropriate to use the median in place of the mean in the LCM model. The approaches based on linear and goal programming presented here examine the two-group classification problems by minimizing the sum of deviations between the classification scores of all objects and the group median scores.

2. Linear programming formulations

Consider the two-group classification problem with k attributes. Let x be the $k \times n$ matrix of attribute scores of the sample with size n drawn from the groups G_1 and G_2. If $w_1, w_2, \ldots, w_k$ are the attribute weights, then the classification score of any object is defined as $S_i = \sum_{j=1}^{k} x_{ij} w_j$. The assignment of an object into a group depends on the value of its classification score. The minimize the sum of deviations model (MSD) can be formulated as follows:

$$(\text{MSD}) \min \sum_{i=1}^{n} d_i \tag{1}$$

s.t.

$$\sum_{j=1}^{k} w_j x_{ij} + d_i \geqslant c, \quad i \in G_1 \tag{2}$$

$$\sum_{j=1}^{k} w_j x_{ij} - d_i \leqslant c, \quad i \in G_2 \tag{3}$$

where $d_i \geqslant 0$ $(i = 1, 2, \ldots, n)$, w_j $(j = 1, 2, \ldots, k)$ and c are unrestricted variables. A normalization constraint is needed to avoid trivial solutions (i.e., an all zero solution). Solving this model gives us the w_j and c values, by which we can obtain the classification score of any object. An object will be classified into G_1 if its classification score is greater than or equal to c, otherwise into G_2.

This model and many of the other existing linear programming models simultaneously determine the attribute weights and cut-off value. Lam et al. (1996) divide the process of their model into two steps: the first constitutes the determination of attribute weights, and the second determines the cut-off value for the classification. In its first step their model makes use of an objective function minimizing the sum of deviations from the group mean classification scores. The LCM model can be formulated as follows:

$$(\text{LCM1}) \min \sum_{i \in G_1, G_2}^{n} (d_i^- + d_i^+) \tag{4}$$

s.t.

$$\sum_{j=1}^{k} w_j (x_{ij} - \mu_{1j}) + d_i^- - d_i^+ = 0, \quad i \in G_1 \tag{5}$$

$$\sum_{j=1}^{k} w_j (x_{ij} - \mu_{2j}) + d_i^- - d_i^+ = 0, \quad i \in G_2 \tag{6}$$

$$\sum_{j=1}^{k} w_j (\mu_{1j} - \mu_{2j}) \geqslant 1 \tag{7}$$