M

Mathematical Programming Models for Classification

Minghe Sun
College of Business, University of Texas at San Antonio, USA

INTRODUCTION

Discriminant and classification analysis has long been fundamental scientific research and practical applications. Discriminant analysis studies the differences between two or more classes of objects represented by measurements, or variables, of different characteristics or attributes. Classification analysis assigns new observations into the classes based on the measurements of the characteristics. Applications of discriminant and classification analysis are diverse. For example, reported applications in business include financial management, human resource management, and marketing; applications in biology and medicine include patient classification, disease diagnosis and species classification; and applications in environment and geography include remote sensing image pattern classification and pollution control.

Based on known values of the attributes or variables and known class memberships of the observations in a sample, classification or discriminant functions are constructed. The attribute values of an observation can be evaluated by these classification or discriminant functions to obtain discriminant scores based on which the observation is assigned to a class. Statistical techniques, such as Fisher's linear discriminant function (Fisher, 1936), Smith's quadratic discriminant function (Smith, 1947) and logistic regression (Hand, 1981), have been standard tools for this purpose. Statistical methods perform well and can provide measures for uncertainty when the data analyzed satisfy the underlying assumptions, such as multivariate normality and equal covariance matrices of the independent variables, although minor deviations from these assumptions do not severely affect the performance of these techniques. More recently, other techniques, such as mathematical programming (MP) (Freed & Glover, 1981a, 1981b; Hand, 1981), neural networks (Stern, 1996) and classification trees (Breiman *et al*., 1984), have become alternative tools to rival statistical techniques. Support vector machine (SVM) (Vapnik, 1995, 1998), as a machine learning technique, is a revolutionary development in classification analysis. SVMs use quadratic programming techniques. Sun (2014) provides an introduction to some SVM models. A spectrum of techniques is needed because no single technique always outperforms others under all situations (Johnson & Wichern, 1988).

MP approaches attracted interests from researchers because these approaches, as nonparametric methods, do not make strict assumptions about the data analyzed, are less influenced by outlier observations and are flexible in incorporating restrictions. The publication of the original linear programming (LP) models for two-class classification (Freed & Glover, 1981a; Hand, 1981) inspired a series of studies. Some of these studies reported pathologies of the earlier MP models, provided diagnoses, and offered remedies (Markowski & Markowski, 1985; Freed & Glover, 1986; Glover *et al*., 1988; Cavalier *et al*., 1989; Glover, 1990; Koehler, 1989a, 1989b, 1990, 1991). The different MP models introduced in the literature include LP, mixed integer programming (MIP), goal programming, nonlinear programming (Erenguc & Koehler, 1990; Stam & Joachimsthaler, 1990; Stam, 1997), multiple objective programming (Stam, 1990; Sun & Xiong, 2003) and quadratic programming (Vapnik, 1995, 1998) models.

DOI: 10.4018/978-1-4666-5202-6.ch136

For decades, research in MP approaches has been focused on the two-class techniques. Formulation of simple but powerful MP models for multi-class discrimination and classification has been an active research topic. A few studies addressed multi-class discrimination and classification with MP approaches and a few MP models have been proposed (Bennett & Mangasarian, 1994; Freed & Glover, 1981b; Gehrlein, 1986; Gochet *et al*., 1997; Sun, 2010). Each of these models has its merits although each may have drawbacks.

Freed and Glover (1981b) proposed a one-function approach and a decomposition approach. In the one function approach, each class is assigned an interval in $\Re$ and an observation is classified into the class where the discriminant score of the observation falls into the interval. In the decomposition approach, a m-class discriminant problem is divided into $m(m-1)/2$ two-class discriminant problems. A model is formulated for each problem representing a pair of classes and is solved separately to determine a classification function representing the hyperplane separating the two classes. Gehrlein (1986) proposed two MIP formulations. Similar to Freed and Glover (1981b), one model uses a single classification function with class specific cut off discriminant scores. The other uses multiple discriminant functions, one for each class. This multiple discriminant function model laid a foundation and also set a limitation for many later studies. The limitation is the use of the difference of two discriminant functions as a constraint. The model proposed by Bennett and Mangasarian (1994) is similar in structure to the multiple discriminant function model proposed by Gehrlein (1986). Instead of minimizing the L_0-norm, Bennett and Mangasarian (1994) minimized a weighted L_1-norm making the model computationally much easier to solve. Gochet *et al*. (1997) proposed a LP model that is also similar in structure to that of Gehrlein (1986) or that of Bennett and Mangasarian (1994). The terms in the objective function of this model are not weighted and a normalization constraint is used.

Some MP models for discriminant and classification analysis are discussed in this chapter, including LP and MIP models. For each of these models, two-class classification and multi-class classification models are discussed. Two examples, one with two and the other with three classes, are used throughout the chapter to demonstrate the formulations of these MP models.

THE CLASSIFICATION PROBLEM

Assume a dataset with m observations in p classes is available for analysis. The index set of the classes is represented by K. The index set of the observations in the whole dataset is represented by I while that in class k is represented by I_k, for $1 \le k \le p$, such that

$$I = \bigcup_{k=1}^{p} I_k .$$

Similarly, the number of observations in class k is represented by m_k such that

$$m = \sum_{k=1}^{p} m_k .$$

Each observation is measured by an input vector of n variables. The variables represent the measurements on the characteristics or attributes of the observations. The input vector for a specific observation $i \in I$ in the dataset is represented by $\mathbf{x}_i \in \Re^n$ and that of a generic observation by $\mathbf{x} \in \Re^n$.

When $p = 2$, a classification function in $\Re^n$ of the form,

$$f(\mathbf{x}) = b_0 + \mathbf{b}'\mathbf{x}, \tag{1}$$

is to be constructed, where $b_0 \in \Re$ and $\mathbf{b} \in \Re^n$ are the estimated parameters. The value of $f(\mathbf{x})$ evaluated at a specific input vector $\mathbf{x}$ is called a classification score. A hyperplane represented by $f(\mathbf{x}) = 0$ is supposed to separate the $p = 2$