Multi-Group Data Classification via MILP

Fadime Üney Yüksektepe
Koç University, Turkey

Metin Türkay
Koç University, Turkey

INTRODUCTION

Data classification is a supervised learning strategy that analyzes the organization and categorization of data in distinct classes. Generally, a training set, in which all objects are already associated with known class labels, is used in classification methods. The data classification algorithms work on this set by using input attributes and builds a model to classify new objects. In other words, the algorithm predicts output attribute values. Output attribute of the developed model is categorical (Roiger & Geatz, 2003). There are many applications of data classification in finance, health care, sports, engineering and science. Data classification is an important problem that has applications in a diverse set of areas ranging from finance to bioinformatics (Chen & Han & Yu, 1996; Edelstein, 2003; Jagota, 2000). Majority data classification methods are developed for classifying data into two groups. As multi-group data classification problems are very common but not widely studied, we focus on developing a new multi-group data classification approach based on mixed-integer linear programming.

BACKGROUND

There are a broad range of methods for data classification problem including Decision Tree Induction, Bayesian Classifier, Neural Networks (NN), Support Vector Machines (SVM) and Mathematical Programming (MP) (Roiger & Geatz, 2003; Jagota, 2000; Adem & Gochet, 2006). A critical review of some of these methods is provided in this section. A major shortcoming of the neural network approach is a lack of explanation of the constructed model. The possibility of obtaining a non-convergent solution due to the wrong choice of initial weights and the possibility of resulting in a non-optimal solution due to the local minima problem are important handicaps of neural network-based methods (Roiger & Geatz, 2003). In recent years, SVM has been considered as one of the most efficient methods for two-group classification problems (Cortes & Vapnik, 1995; Vapnik, 1998). SVM method has two important drawbacks in multi-group classification problems; a combination of SVM has to be used in order to solve the multi-group classification problems and some approximation algorithms are used in order to reduce the computational time for SVM while learning the large scale of data.

There have been numerous attempts to solve classification problems using mathematical programming (Joachimsthaler & Stam, 1990). The mathematical programming approach to data classification was first introduced in early 1980's. Since then, numerous mathematical programming models have appeared in the literature (Erenguc & Koehler, 1990) and many distinct mathematical programming methods with different objective functions are developed in the literature. Most of these methods modeled data classification as linear programming (LP) problems to optimize a distance function. In addition to LP problems, mixed-integer linear programming (MILP) problems that minimize the misclassifications on the design data set are also widely studied. There have been several attempts to formulate data classification problems as MILP problems (Bajgier & Hill, 1982; Gehrlein 1986; Littschwager, 1978; Stam & Joachimsthaler, 1990). Since MILP methods suffer from computational difficulties, the efforts are mainly focused on efficient solutions for two-group supervised classification problems. Although it is possible to solve a multi-group data classification problem by solving several two-group problems, such approaches also have drawbacks including computational complexity resulting in long computational times (Tax & Duin, 2002).

MAIN FOCUS

The objective in data classification is to assign data points that are described by several attributes into a predefined number of groups. The use of hyper-boxes for defining boundaries of the sets that include all or some of the points in that set as shown in Figure 1 can be very accurate for multi-group problems. Hyper-boxes are high dimensional geometrical shapes that have lower and upper boundaries for each attribute. If it is necessary, more than one hyper-box can be used in order to represent a group as shown in Figure 1.

The data classification problem based on this idea is developed in two parts: training and testing. During the training part, characteristics of data points that belong to a certain group are determined and differentiated from the data points that belong to other groups. After the distinguishing characteristics of the groups are determined, then the effectiveness of the classification must be tested. Predictive accuracy of the developed model is performed on a test data set during the testing stage.

Training Problem Formulation

Training is performed on a training data set composed of a number of instances i. The data points are represented by the parameter a_{im} that denotes the value of attribute m for the instance i. The group k that the data point i belongs to are given by the set D_{ik}. Each existing hyper-box l encloses a number of data points belonging to group k. Moreover, bounds n (*lower, upper)* of each hyper-box is determined by solving the training problem.

Given these parameters and the sets, the following variables are sufficient to model the multi-group data classification problem with hyper-boxes. The binary variable yb_l is indicates whether the box l is used or not. The position (inside or outside) of the data point i with regard to box l is represented by ypb_{il}. The assigned group k of box l and data point i is symbolized by ybc_{lk} and ypc_{ik}, respectively. If the data point i is within the bound n with respect to attribute m of box l, then the binary variable $ypbn_{ilmn}$ takes the value of 1, otherwise 0. Similarly, $ypbm_{ilm}$ indicates whether the data point i is within the bounds of attribute m of box l or not. Finally, yp_{ik} indicate the misclassification of data point i to group k. In order to define the boundaries of hyper-boxes, two continuous variables are required: X_{lmn} is the one that models bounds n for box l on attribute m. Correspondingly, bounds n for box l of group k on attribute m are defined with the continuous variable $XD_{l,k,m,n}$.

The following MILP problem models the training part of multi-group data classification method using hyper-boxes:

$$\min \; z = \sum_{l}\sum_{k} yp_{ik} + \sum_{l} yb_l \qquad (1)$$

subject to

Figure 1. Schematic representation of multi-group data classification using hyper-boxes.