Chapter II
Vector DNF for Datasets Classifications:
Application to the Financial Timing Decision Problem

Massimo Liquori
Università di Roma "La Sapienza", Italy

Andrea Scozzari
Università di Roma "La Sapienza", Italy

ABSTRACT

Traditional classification approaches consider a dataset formed by an archive of observations classified as positive or negative according to a binary classification rule. In this chapter, we consider the financial timing decision problem, which is the problem of deciding the time when it is profitable for the investor to buy shares or to sell shares or to wait in the stock exchange market. The decision is based on classifying a dataset of observations, represented by a vector containing the values of some financial numerical attributes, according to a ternary classification rule. We propose a new technique based on partially defined vector Boolean functions. We test our technique on different time series of the Mibtel stock exchange market in Italy, and we show that it provides a high classification accuracy, as well as wide applicability for other classification problems where a classification in three or more classes is needed.

INTRODUCTION

In the area of knowledge-based expert systems, the aim is to detect structural information from large datasets in order to extract salient features for identifying differences that separate one set of data from another. Classification methods developed in the literature try to classify the given observations

and, in addition, to classify new observations in a way consistent with past classifications. Such structural information can provide powerful means for the solution of a variety of problems, including classification, automated knowledge acquisition for expert systems, development of pattern-based decision support systems, detection of inconsistencies in databases, feature selection, medical diagnosis, marketing, and numerous aspects of etiology.

Several approaches coming from different fields have been proposed in the literature to tackle the classification problem. One of the best-known methods is support vector machines (SVM) that has proved highly successful in a number of classification studies. Although the subject traces its origin to the seminal work of Vapnik and Lerner (1963), it is only now receiving a growing attention. In the simplest case, given a set of observations classified into two classes, the aim is to construct a function to discriminate between classes. This can be done via a mathematical programming approach. A linear programming-based approach, stemming from the multisurface method of Mangasarian (1965, 1968), has been used for a breast cancer diagnosis system (Mangasarian, Setiono, & Wolberg, 1990; Mangasarian,, Street, & Wolberg, 1995; Wolberg, & Mangasarian, 1990). Another approach is the quadratic programming method based on Vapnik's Statistical Learning Theory (Cortes & Vapnik, 1995; Vapnik1995). See Burges (1998) for a tutorial on classification via SVMs. Bredensteiner and Bennett (1999) show how the linear programming and quadratic programming methods can be combined to yield two new approaches for the multiclass problem. Other mathematical programming techniques, based on the minimization of some function measuring the classification error (Freed & Glover, 1986; Glover, 1990; Kamath, Karmarkar, Ramakrishnan, & Resende, 1992; Triantaphyllou, Allen, Soyster, & Kumara, 1994), have been used in classification problems. A MINSAT approach for learning logic relationship that correctly classify a given dataset has been recently proposed in Felici and Truemper (2000).

Decision trees are another popular technique for classification. The main reason behind their popularity seems to be their relative advantage in terms of interpretability. There are several efficient and simple implementations of decision trees (Quinlan, 1993). In a recent work, Street (2004) presents an algorithm based on nonlinear programming for multicategory decision trees. Unfortunately, one of the limitations of most decision trees is that they are known to be unstable, especially when dealing with large data sets (Fu, Golden, Lele, Raghavan, & Wasil, 2003). In the literature, there are several papers that provide heuristics and metaheuristics for the problem of finding an optimal decision tree, which is known to be an NP-complete problem (Fu, et al., 2003; Niimi & Tazaki, 2000).

Naive Bayes method is another simple but effective classifier (Jefferys & Berger, 1992; Yeung, 1993). The attributes, observed in the training set, are assumed to be conditionally independent, given the value of the class attribute. In order to derive a good classification rule, and considering the independence assumption made, the marginal probabilities of each attribute must be estimated. In Lin (2002), it is shown that the asymptotic target of support vector machines is some interesting classification functions that are directly related to the Bayes rule. Actually, the independence assumption is unrealistic, thus, Bayesian networks have been introduced that explicitly model dependencies between attributes (Pernkopf, 2005). Thus, given a set of observations, the problem is to find a network that best matches the training set. The search for the best network is based on a scoring function that evaluates each network with respect to the training data (Heckerman, Geiger, & Chickering, 1995; Lam & Bacchus,1994).

Several classification problems can also be formulated as an artificial neural network problem. An artificial neural network (ANN) can be thought of as a mathematical paradigm that models the bio-