Computers & Operations Research 40 (2013) 150–165

Contents lists available at SciVerse ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Review

Supervised classification and mathematical optimization ☆

Emilio Carrizosa [a], Dolores Romero Morales [b,*]

[a] *Universidad de Sevilla, Spain*
[b] *University of Oxford, United Kingdom*

ARTICLE INFO

Available online 13 June 2012

Keywords:
Data mining
Mathematical optimization
Support vector machines
Interpretability
Cost efficiency

ABSTRACT

Data mining techniques often ask for the resolution of optimization problems. Supervised classification, and, in particular, support vector machines, can be seen as a paradigmatic instance. In this paper, some links between mathematical optimization methods and supervised classification are emphasized. It is shown that many different areas of mathematical optimization play a central role in off-the-shelf supervised classification methods. Moreover, mathematical optimization turns out to be extremely useful to address important issues in classification, such as identifying relevant variables, improving the interpretability of classifiers or dealing with vagueness/noise in the data.

Contents

☆ This work has been partially supported by projects MTM2009-14039 and FQM-329 of Junta de Andalucía, Spain.

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), dolores.romero-morales@sbs.ox.ac.uk (D. Romero Morales).

0305-0548/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2012.05.015

1. Introduction

One of the most important tasks in data mining [116,121,240] is *supervised classification*, which seeks procedures for classifying objects in a set Ω into a set $\mathcal{C}$ of classes. Each object $u \in \Omega$ has associated a pair $(\mathbf{x}^u, y^u)$, where $\mathbf{x}^u$, the *predictor vector*, takes values on a set X, usually assumed to be a subset of $\mathbb{R}^p$, and $y^u \in \mathcal{C}$ is the class membership of u. Hereafter, we will simply use the term *variable* to refer to each component of the predictor vector.

Not all the information about the objects in Ω is available: the class membership c^u is only known for those objects u in some subset $I \subset \Omega$, called the *training sample*.

With this information, a classification rule is sought, i.e., a function $y : X \to \mathcal{C}$, which assigns label $y(\mathbf{x}) \in \mathcal{C}$ to predictor vector $\mathbf{x}$, $\forall \mathbf{x}$.

In its basic form, $\mathcal{C}$ consists of a finite set of nominal values, without an intrinsic ranking (e.g. $\mathcal{C} = \{\text{benign, malign}\}$), though part of the theory extends to the case in which $\mathcal{C}$ is a finite set equipped with an order relation, or a segment of the real line. In this latter case, we would have a regression problem instead [216].

Supervised classification has been successfully applied in many different fields. Examples are found in text categorization [210], such as document indexing, webpage classification and spam filtering; biology and medicine, such as classification of gene expression data [102,245], homology detection [143], protein–protein interaction prediction [28,171], abnormal brain activity classification [55] and cancer diagnosis [110,165]; machine vision [75,188]; agriculture [179]; or chemistry [60], to cite a few fields and references.

We can argue that business applications have had a later start. Despite of this, nowadays we can find many applications of supervised classification in marketing, customer relationship management, banking, among others [4,5,115,144,153,175]. Typical business applications are credit scoring [11], bankruptcy [176], fraud detection [87], customer targeting [73], customer loyalty [106,112], market basket analysis [61], recommender systems [62], revenue management [122], services booking cancellations [203], country risk ratings [114], prediction of health costs [24] or stock market forecast [103,156].

Mathematical optimization has played a crucial role in supervised classification [21,22,31,32,84,81,88,107,218,242]. Techniques from very diverse fields within mathematical optimization have been shown to be useful. As we will discuss in Section 3, and already pointed out e.g. in [17], many of the optimization problems encountered fall within the area of (smooth) Convex Programming. However, other areas of mathematical optimization play a notable role, among others, global optimization [9,13,51,128,160,245], linear programming [94,158,205] mixed-integer programming [25,39,50,77,220,228], nonsmooth optimization [7,13,44,51,222,223], multicriteria and multi-objective programming [68,93,181,248] and robust optimization [224].

The success of mathematical optimization when applied to supervised classification has one of its main exponents in Support Vector Machines (SVMs) [30,72,229,230], a technique rooted in statistical learning theory [229,230], which has proved to be one of the state-of-the-art methods for supervised learning. For the two-class case, SVM aims at separating both classes by means of a hyperplane which maximizes the margin, i.e., the width of the band separating the two sets. This geometrical optimization problem can be written as a convex quadratic optimization problem with linear constraints, in principle solvable by any nonlinear optimization procedure. See also [18,40,108,177,113] for introductory surveys on SVMs.

The purpose of this paper is to illustrate that mathematical optimization is at the core of supervised classification methods. Moreover, the variety of algorithmic tools which have been used is rather wide, implying that the researchers in different branches of mathematical optimization have ample room to translate their expertise into this context, and they may find new domains of applicability to existing mathematical optimization knowledge. We stress that the aim of this paper is not to study in depth classification methods, and we refer the reader to [36,121,125,197,201] for further details. Instead, our aim is to illustrate the central role played by mathematical optimization in such methods.

Due to its performance and its optimization context, we have a special focus on SVMs. The two basic versions of SVM, namely, the hard and soft margin approaches, are introduced. We then move on to the embedding of the variable space into a feature space of higher dimension, and the kernel version of SVM. We discuss the optimization problem behind the SVM, as well as the mathematical optimization techniques that have been proposed to solve it. We devote the rest of the paper to extensions of SVM dealing with critical issues such as interpretability, cost efficiency or robustness, as well as dealing with data that may be unbalanced, imprecise or unlabeled. We will show that many different domains in mathematical optimization are needed to cope with both the standard SVM formulation as well as the variants addressing the properties mentioned above.

The remainder of the paper is organized as follows. In Section 2 we briefly discuss off-the-shelf classification methods and the role of mathematical optimization. In Section 3 we introduce the SVM, including the hard and soft versions as well as its kernelization. In Section 4 we discuss the mathematical optimization techniques proposed in the literature to build SVMs. In Section 5 we discuss critical modeling issues and how mathematical optimization can help to incorporate them into the original SVM model. Finally, conclusions are provided in Section 6.

2. Classification methods

Different classification methods have been proposed in the literature. They mainly differ in the statistical assumptions made of the data and the type of algorithms needed to construct the classifier. In this section we review benchmarking classification methods divided into three main categories: linear classifiers, nearest-neighbor classifiers and classification trees. The methods proposed have a deep geometrical flavor. For data sets in very low dimension, computational geometry tools may be useful [23]. However, most real world data sets of interest are in larger dimension, calling for numerical rather than geometrical procedures.

Before presenting the benchmarking classification methods, we briefly discuss two important ingredients in supervised learning: the scoring functions, a general framework used to describe the classifier, and the performance criteria used to compare classification methods.

2.1. *Scoring functions*

The methods reviewed below are based on scoring functions: for each $c \in \mathcal{C}$, a scoring function $f_c : X \to \mathbb{R}$ is built from the training set I, and forthcoming objects with predictor vector $\mathbf{x}$ are classified as members of the class $y(\mathbf{x})$ with highest score. In other words, the classifier is given by the function

$$y(\mathbf{x}) \in \arg\max_{c \in \mathcal{C}} f_c(\mathbf{x}). \tag{1}$$

In case of ties, objects are randomly assigned to one of the classes at which the maximum is attained.

The scoring function f_c ranks the objects, so that, the higher the value of $f_c(\mathbf{x})$, the higher the likelihood that an object represented by $\mathbf{x}$ is in class c.