Logical analysis of numerical data

Endre Boros a,2, Peter L. Hammer a,*, Toshihide Ibaraki b,3,
Alexander Kogan a,c,4

a RUTCOR, Rutgers University, P.O. Box 5062, New Brunswick, NJ 08903, USA
b Department of Applied Mathematics and Physics, Graduate School of Engineering,
Kyoto University, 606 Kyoto, Japan
c Department of Accounting and Information Systems, Faculty of Management, Rutgers University,
Newark, NJ 07102, USA

Received 27 February 1997; accepted 21 April 1997

Abstract

"Logical analysis of data" (LAD) is a methodology developed since the late eighties, aimed at
discovering hidden structural information in data sets. LAD was originally developed for analyzing
binary data by using the theory of partially defined Boolean functions. An extension of LAD for
the analysis of numerical data sets is achieved through the process of "binarization" consisting in
the replacement of each numerical variable by binary "indicator" variables, each showing whether
the value of the original variable is above or below a certain level. Binarization was successfully
applied to the analysis of a variety of real life data sets. This paper develops the theoretical
foundations of the binarization process studying the combinatorial optimization problems related
to the minimization of the number of binary variables. To provide an algorithmic framework
for the practical solution of such problems, we construct compact linear integer programming
formulations of them. We develop polynomial time algorithms for some of these minimization
problems, and prove NP-hardness of others. © 1997 The Mathematical Programming Society,
Inc. Published by Elsevier Science B.V.

Keywords: Data analysis; Boolean functions; Machine learning; Binarization; Set covering; Monotonicity;
Thresholdness; Computational complexity
1. Introduction

1.1. Logical analysis of data

The study of numerical data, common to a large variety of disciplines, leads to interesting problems of combinatorial optimization, the solution of which can have an immediate impact on the understanding of the nature of the phenomena under investigation, the factors governing them, and on the ways we can influence their development.

The data we are examining here consists of collections of "observations" represented as points in \( \mathbb{R}^d \). These observations can be associated to patients in a medical study, consumers in a marketing analysis, probes in a potential oil field, etc. The components of the vectors, called "attributes" stand for the results of various medical tests, financial parameters, geological measurements, etc. The observations fall in two classes: "positive" ones (e.g. healthy patients, potential buyers, oil rich areas, etc.) and "negative" ones (e.g. sick patients, etc.).

The goal of data analysis is to arrive to logical explanations distinguishing positive and negative observations. It was shown in [11,15] that - in the case of binary data - this goal can be achieved by the use of partially defined Boolean functions.

The usefulness of logical or Boolean techniques is demonstrated by many approaches in this field [1,10,12,17,22,23]. The classification power of the Boolean-based methodology of the logical analysis of data (LAD) and its applicability to oncology, psychiatry, oil exploration, economic analysis, and other fields are described in [7,16,19]. Preliminary results reported in [7] already indicate that the LAD approach not only has high classification power for distinguishing between positive and negative examples, but also is quite successful in extracting underlying structural information from the given data, which is useful in understanding why and how phenomena occur. While the techniques of LAD are entirely different from those of the pioneering works of Mangasarian (e.g. [18]), the two approaches provide strongly complementary and mutually reinforcing results.

The methodology of LAD is extended to the case of numerical data by a process called binarization, consisting in the transformation of numerical (real valued) data to binary (0,1) ones. In this transformation we map each observation \( u = (u_A, u_B, \ldots) \) of the given numerical data set to a binary vector \( x(u) = (x_1, x_2, \ldots) \in \{0,1\}^n \) by defining e.g. \( x_1 = 1 \) iff \( u_A \geq \alpha_1 \), \( x_2 = 1 \) iff \( u_B \geq \alpha_2 \), etc., and in such a way that if \( u \) and \( v \) represent a positive and a negative observation point respectively, then \( x(u) \neq x(v) \). The binary variables \( x_i, i = 1, \ldots, n \) associated to the real attributes are called indicator variables, and the real parameters \( \alpha_i, i = 1, \ldots, n \) used in the above process are called cut points.

The binarization process described above has been included in a current implementation of LAD and successfully used in the computational experiments reported in [7].

The object of this paper is to provide a theoretical foundation for the binarization process with the primary focus on the combinatorial optimization problems related to the minimization of the number of cut points. The paper studies the computational