Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 346 (2019) 65–76

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

The Geodesic Classification Problem on Graphs

Paulo Henrique Macêdo de Araújo [1,3]

Departamento de Computação and Campus de Quixadá
Universidade Federal do Ceará
Fortaleza and Quixadá, CE, Brazil

Manoel Campêlo[2,4]

Departamento de Matemática e Estatística Aplicada
Universidade Federal do Ceará
Fortaleza, CE, Brazil

Ricardo C. Corrêa[2,5]

Departamento de Ciência da Computação
Universidade Federal Rural do Rio de Janeiro
Nova Iguaçu, RJ, Brazil

Martine Labbé [6]

Departement D'Informatique
Université Libre de Bruxelles
Bruxelles, Belgium

Abstract

Motivated by the significant advances in integer optimization in the past decade, Bertsimas and Shioda developed an integer optimization method to the classical statistical problem of classification in a multidimensional space, delivering a software package called *CRIO* (*Classification and Regression via Integer Optimization*). Following those ideas, we define a new classification problem, exploring its combinatorial aspects. That problem is defined on graphs using the geodesic convexity as an analogy of the Euclidean convexity in the multidimensional space. We denote such a problem by *Geodesic Classification* (*GC*) problem. We propose an integer programming formulation for the *GC* problem along with a branch-and-cut algorithm to solve it. Finally, we show computational experiments in order to evaluate the combinatorial optimization efficiency and classification accuracy of the proposed approach.

Keywords: Classification, Geodesic Convexity, Integer Linear Programming.

https://doi.org/10.1016/j.entcs.2019.08.007

1 Introduction

Supervised learning denotes the automatic prediction of the behavior of unknown data based on a set of samples. It is a tool widely used in many everyday situations of the information society in which we live. In general terms, it can be described as the following two-phase procedure: in the initial phase, or *training phase*, the sample set is analyzed. Each sample consists of an array of encoded attributes that characterize an object of a certain type together with a label that associates a class to the corresponding object. Most commonly, only two classes are considered. A tacit assumption made at this phase is that there is an underlying pattern associated with the samples of each class that sets them apart from the samples of the other classes. Thus, the purpose of the training phase is to determine a mapping from all possible objects into the set of possible classes as an extension of an underlying patterns of the samples. Then, in the second phase, the mapping determined in the training phase is used to respond to queries for the class of objects that do not belong to the sample set.

An optimization problem is usually associated with the training phase. Referred to as *classification problem*, it consists in grouping similar samples so as to get clusters as internally homogeneous as possible. A wide range of solution methods is available, each depending on the coding of the samples and the criterion adopted to express homogeneity. A very popular approach is to encode the samples as vectors in an Euclidean space and to assume that the class patterns can be appropriately characterized by convex sets. In this vein, continuous optimization methods, including linear and quadratic programming, have been developed in the last 40 years [1,6,8,9,14]. More recently, integer linear programming tools started to be used in conjunction with continuous methods [3,13,16,17,18].

Strongly inspired by the version of the classification problem based on Euclidean convexity concepts discussed in [7], the object of study of this paper is the use of integer linear programming formulations for the resolution of a new variation of the classification problem stated in terms of notions of convexity in graphs. The statement of the new classification problem assumes the following hypotheses:

(i) The objects are not encoded numerically. Instead, each object is characterized by its similarities with other objects. The configuration of the objects is thus represented by a similarity graph $G = (V, E)$, connected, where V is the set of all objects and E gives the pairs of similar objects. The objects associated with the sample set constitute a proper subset of V. In addition, it is assumed that there is an underlying samples pattern that can be expressed, or at least

[1] Partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

[2] Partially supported by the Brazilian agencies CAPES, CNPq (Proc. 305264/2016-8, Proc. 443747/2014-8) and FUNCAP (PNE-0112-00061.01.00/16).

[3] Email: phmacedoaraujo@lia.ufc.br

[4] Email: mcampelo@lia.ufc.br

[5] Email: correa@ufrrj.br

[6] Email: mlabbe@ulb.ac.be