J. Algorithms 64 (2009) 79–88

Contents lists available at ScienceDirect

Journal of Algorithms
Cognition, Informatics and Logic

www.elsevier.com/locate/jalgor

Bichromatic separability with two boxes: A general approach

C. Cortés [a], J.M. Díaz-Báñez [b,*,1], P. Pérez-Lantero [c,2], C. Seara [d,3], J. Urrutia [e,1], I. Ventura [b,1]

[a] Departamento Matemática Aplicada I, Universidad de Sevilla, Spain
[b] Departamento Matemática Aplicada II, Universidad de Sevilla, Spain
[c] Departamento de Computación, Universidad de La Habana, Cuba
[d] Departament de Matemàtica Aplicada II, Universitat Politècnica de Catalunya, Spain
[e] Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico

ARTICLE INFO

Article history:
Received 16 September 2008
Available online 12 February 2009

Keywords:
Discrete optimization
Maximum consecutive subsequence
Dynamic maintenance
Algorithm design
Classification problems
Bioinformatics

ABSTRACT

Let S be a set of n points on the plane in general position such that its elements are colored red or blue. We study the following problem: *Find a largest subset of S which can be enclosed by the union of two, not necessarily disjoint, axis-aligned rectangles $\mathcal{R}$ and $\mathcal{B}$ such that $\mathcal{R}$ (resp. $\mathcal{B}$) contains only red (resp. blue) points.* We prove that this problem can be solved in $O(n^2 \log n)$ time and $O(n)$ space. Our approach is based on solving some instances of Bentley's maximum-sum consecutive subsequence problem. We introduce the first known data structure to dynamically maintain the optimal solution of this problem. We show that our techniques can be used to efficiently solve a more general class of problems in data analysis.

1. Introduction

In data mining and classification problems, a natural method for analyzing data is to select prototypes representing different data classes. A standard technique for achieving this is to perform cluster analysis on the training data [8,11]. The clusters can be obtained using simple geometric shapes such as circles or boxes. Aronov and Har-Peled [1] and Eckstein et al. [9] considered circles and axis-aligned boxes for the selection problem. Aronov and Har-Peled [1] studied the following problem: *Given a bicolored point set, find a ball that contains the maximum number of red points without containing any blue points*. This type of classification is *asymmetric* in the way it treats red and blue points. The goal is to separate the "red class" from the "blue class". We are interested in generalizing to a *symmetric two-classes problem* by finding a witness set for each color.

In some cases, as in medical data analysis [10], methods can produce biased classifications due to the fact that some data may be defective or contain values out of reasonable ranges. In other cases, we may obtain data which are hard to classify due to strong similarities between subsets of different classes. A potential way to find a better classification in the former problem is to remove some data points from the input. Culling the minimum number of such points can be a suitable criterion for losing the smallest amount of information possible. We will use the following notation. Given two (possibly unbounded) non-disjoint convex sets X and Y, we denote by $|X|$ the cardinality of the set X and by $X \setminus Y$ the set obtained by removing $X \cap Y$ from X.

* Corresponding author at: Departamento Matemática Aplicada II, Universidad de Sevilla, Escuela Superior de Ingenieros, Avda. de los Descubrimientos, s/n, 41092, Sevilla, Spain. Fax: +34 95 4486165.
E-mail addresses: ccortes@us.es (C. Cortés), dbanez@us.es (J.M. Díaz-Báñez), pablo@matcom.uh.cu (P. Pérez-Lantero), carlos.seara@upc.edu (C. Seara), urrutia@matem.unam.mx (J. Urrutia), iventura@us.es (I. Ventura).

[1] Partially supported by Grant MEC MTM2006-03909.
[2] Partially supported by Grant MAEC-AECI and MEC MTM2006-03909.
[3] Partially supported by Grants MEC MTM2006-01267 and DURSI 2005SGR00692.

0196-6774/$ – see front matter
doi:10.1016/j.jalgor.2009.01.001

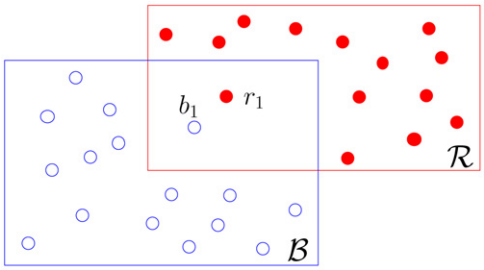

Fig. 1. Getting a solution by removing the points r_1 and b_1.

In this paper we study the following problem:

The Two Enclosing Boxes problem (*2-EB-problem*). Let S be a set of n points on the plane in general position such that the points are colored red or blue. Compute two open axis-aligned rectangles $\mathcal{R}$ and $\mathcal{B}$ such that the number of red points in $\mathcal{R} \setminus \mathcal{B}$ plus the number of blue points in $\mathcal{B} \setminus \mathcal{R}$ is maximized.

We remark here that in the definition of our problem we require $\mathcal{R}$ and $\mathcal{B}$ to be open. This will facilitate our presentation, but in practice we may proceed in a similar way if our boxes are closed. Observe that $\mathcal{R}$ and $\mathcal{B}$ may intersect; however any point in $\mathcal{R} \cap \mathcal{B}$ has to be removed from S. For example, the solution to the 2-EB-problem for the point set S illustrated in Fig. 1 is $n-2$, where n is the cardinality of the input set. By removing r_1 and b_1 from S, we can obtain two rectangles $\mathcal{R}$ and $\mathcal{B}$, each of them containing only red and blue points respectively. Notice that an asymmetric separation approach as the one used by Aronov and Har-Peled [1] does not give a solution to our problem, so we must design a procedure which consider $\mathcal{R}$ and $\mathcal{B}$ simultaneously. Bespamyatnikh and Segal [4] studied a two-box covering problem but using a different min–max criterion.

The 2-EB-problem was first introduced by Cortés et al. [5], solving the problem with an $O(n^3)$-time and space algorithm. In this paper we show that the 2-EB-problem can be solved in $O(n^2 \log n)$ time and $O(n)$ space. We also introduce a new data structure that allows us to dynamically solve Bentley's [2] well known *Maximum-Sum Consecutive Subsequence* problem (MCS-problem for short) together with some other variants of this problem that have applications, for instance in sequence analysis in bioinformatics [6]. We also show a generalization of our approach that can be used to solve a general type of problem of interest in areas such as computer graphics or machine learning [7].

The outline of this paper is as follows. In Section 2 we introduce a data structure, the MCS-tree, to dynamically maintain an optimal solution of the MCS-problem. In Section 3 we introduce some notation and present the first results on the 2-EB-problem. In Section 4 we show our main result, an $O(n^2 \log n)$ time and linear space algorithm to solve the 2-EB-problem. In Section 5 we show how to solve the following related problem: *Let S be a set of points on the plane in general position such that each element of S is colored red, blue, or green. Find three pairwise-disjoint axis-aligned rectangles $\mathcal{R}$, $\mathcal{B}$, and $\mathcal{G}$ such that the total number of red, blue, and green points contained in $\mathcal{R}$, $\mathcal{B}$, and $\mathcal{G}$ respectively is maximized.* In Section 6 we present a generalization of our technique and show how to apply it to several variants of the original problem. Finally, in Section 7 we present the conclusions.

2. The dynamic MCS-problem

In this section we describe the main tool that will allow us to solve the 2-EB-problem in $O(n^2 \log n)$ time and $O(n)$ space. Later we show how this technique can be applied to other variants of the same problem. The key idea for solving the 2-EB-problem is a reduction to the computation of some dynamic instances of the following one-dimensional Bentley's problem [2]:

The Maximum-Sum Consecutive Subsequence problem *(MCS-problem).* Given a sequence $X = (x_1, x_2, \ldots, x_n)$ and a real weight function w over its elements where for each i, $w(x_i)$ is not necessarily positive, compute the consecutive subsequence $(x_i, x_{i+1}, \ldots, x_j)$ of X such that $w(x_i) + w(x_{i+1}) + \cdots + w(x_j)$ is maximum.

Next we show how to construct a binary tree, the MCS-tree, that allows us to solve the MCS-problem in a dynamic way. Assume that n is a power of two, otherwise add a few elements with negative weights at the end of the sequence $X = (x_1, \ldots, x_n)$ until we get a sequence of 2^k elements, $n < 2^k < 2n$.

2.1. The MCS-tree

The MCS-tree is a balanced binary tree with n leaves representing the sequence $X = (x_1, x_2, \ldots, x_n)$. The kth leaf (from left to right) represents x_k. Each internal node u represents the consecutive subsequence formed by the descendants (leaves)