Fitting a Step Function to a Point Set

Hervé Fournier, Antoine Vigneron

To cite this version:

Hervé Fournier, Antoine Vigneron. Fitting a Step Function to a Point Set. 16th Annual European Symposium on Algorithms, Sep 2008, Karlsruhe, Germany. 10.1007/978-3-540-87744-8_37 . hal-02750559

HAL Id: hal-02750559
https://hal.inrae.fr/hal-02750559

Submitted on 3 Jun 2020

Fitting a Step Function to a Point Set

Hervé Fournier[1] and Antoine Vigneron[2]

[1] Laboratoire PRiSM
CNRS UMR 8144 and Université de Versailles St-Quentin en Yvelines
45 av. des États-Unis, 78035 Versailles, France
herve.fournier@prism.uvsq.fr
[2] INRA, UR 341 Mathématiques et Informatique Appliquées
78352 Jouy-en-Josas, France
antoine.vigneron@jouy.inra.fr

Abstract. We consider the problem of fitting a step function to a set of points. More precisely, given an integer k and a set P of n points in the plane, our goal is to find a step function f with k steps that minimizes the maximum vertical distance between f and all the points in P. We first give an optimal $\Theta(n \log n)$ algorithm for the general case. In the special case where the points in P are given in sorted order according to their x-coordinates, we give an optimal $\Theta(n)$ time algorithm. Then, we show how to solve the weighted version of this problem in time $O(n \log^4 n)$. Finally, we give an $O(nh^2 \log h)$ algorithm for the case where h outliers are allowed, and the input is sorted. The running time of all our algorithms is independent of k.

1 Introduction

In this paper, we consider the problem of fitting a step function to a point set in $\mathbb{R}^2$. (See Figure 1 for an example.) For a given number of steps k, we find the step function whose maximum vertical distance to a point set P is minimized.

The motivation for this work is to find a concise representation of a large set of points by a step-function with few steps. This type of representation of point-sets by step-functions (also called *histograms*) is used in Database Management Systems, for *query optimization*: there can be many different ways of answering a complex query involving several attributes of the data, and query optimization consists in predicting the fastest way to answer a query, before this query is performed. This prediction is done using some statistics on the data, which usually consists of histograms. Several types of histograms have been used in databases [13]; the histograms that correspond to our optimal step-functions are called *maximum error histograms*, and have recently been studied in the database community [4,10,14].

We give optimal algorithms for computing the optimal step-function in our model; in other words, we give optimal algorithms for computing maximum error histograms. We also give efficient algorithms for two generalizations. First, we consider the more general case where each point is weighted, and its contribution to the distance computation is multiplied by this weight. Second, we introduce a generalization of our problem to the case where outliers are allowed; that is, we allow our algorithm to ignore h input

D. Halperin and K. Mehlhorn (Eds.): ESA 2008, LNCS 5193, pp. 442–453, 2008.