Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation.

Antonio Irpino*, Elvira Romano**

* Dipartimento di studi europei e mediterranei
Seconda Università degli Studi di Napoli
Via del Setificio, 15 Complesso Monumentale Belvedere - San Leucio
I-81020 Caserta (CE)
irpino@unina.it
** Dipartimento di Matematica e Statistica
Universita' degli Studi di Napoli "Federico II"
Via Cintia - Complesso Monte Sant'Angelo
I-80126 Napoli
elvrom@unina.it

Summary. Histogram representation of a large set of data is a good way to summarize and visualize data and is frequently performed in order to optimize query estimation in DBMS. In this paper, we show the performance and the properties of two strategies for an optimal construction of histograms on a single real valued descriptor on the base of a prior choice of the number of buckets. The first one is based on the Fisher algorithm, while the second one is based on a geometrical procedure for the interpolation of the empirical distribution function by a piecewise linear function. The goodness of fit is computed using the Wasserstein metric between distributions. We compare the proposed method performances against some existing ones on artificial and real datasets.

1 Introduction

Today's storage information mechanism fails to capture a large amount of data and preprocess them in their entirety, while only a summary is stored. In this context *histogram* plays the role of a tool for producing a suitable summarizing description and quickly answering to decision support queries. Following the guide phrase *"An image says more than one hundred words"*, the histogram represents a simple and intuitive graphical tool to describe data distribution. It smoothes the data to display the general shape of an empirical distribution. The problem is that it can give a false impression of the shape of the dataset distribution, because its construction depends on the choice of the number and the length of the subintervals - usually called *buckets or bins* - of the real lines on which the histogram is based. Ideally it could have the situation in which for large bins the nature of the dataset is bimodal and for small bins the plot reduces to unimodal representation. The matter at stake here concerns the *kind of bin width that can take into account the best graphical representation of the underlying DBMS and how it can be constructed with minimal error approximation.*