Selectivity estimators for multidimensional range queries over real attributes

Dimitrios Gunopulos\textsuperscript{1}, George Kollios\textsuperscript{2}, Vassilis J. Tsotras\textsuperscript{3\textsuperscript{1}}, Carlotta Domeniconi\textsuperscript{3}

\textsuperscript{1} University of California, Riverside, Department of Computer Science and Engineering, Bourns College of Engineering, Riverside, CA 92521, USA; e-mail: \{dg, tsortas\}@cs.ucr.edu
\textsuperscript{2} Boston University, Department of Computer Science, Boston, MA 02215, USA; e-mail: gkollios@cs.bu.edu
\textsuperscript{3} George Mason University, Department of Information and Software Engineering, Fairfax, VA 22030, USA; e-mail: carlotta@ise.gmu.edu

Edited by Y. Ioannidis. Received: January 30, 2001 / Accepted: June 9, 2003
Published online: March 4, 2004 – © Springer-Verlag 2004

Abstract. Estimating the selectivity of multidimensional range queries over real valued attributes has significant applications in data exploration and database query optimization. In this paper, we consider the following problem: given a table of \(d\) attributes whose domain is the real numbers and a query that specifies a range in each dimension, find a good approximation of the number of records in the table that satisfy the query. The simplest approach to tackle this problem is to assume that the attributes are independent. More accurate estimators try to capture the joint data distribution of the attributes. In databases, such estimators include the construction of multidimensional histograms, random sampling, or the wavelet transform. In statistics, kernel estimation techniques are being used. Many traditional approaches assume that attribute values come from discrete, finite domains, where different values have high frequencies. However, for many novel applications (as in temporal, spatial, and multimedia databases) attribute values come from the infinite domain of real numbers. Consequently, each value appears very infrequently, a characteristic that affects the behavior and effectiveness of the estimator. Moreover, real-life data exhibit attribute correlations that also affect the estimator. We present a new histogram technique that is designed to approximate the density of multidimensional datasets with real attributes. Our technique defines buckets of variable size and allows the buckets to overlap. The size of the cells is based on the local density of the data. The use of overlapping buckets allows a more compact approximation of the data distribution. We also show how to generalize kernel density estimators and how to apply them to the multidimensional query approximation problem. Finally, we compare the accuracy of the proposed techniques with existing techniques using real and synthetic datasets. The experimental results show that the proposed techniques behave more accurately in high dimensionalities than previous approaches.

1 Introduction

Computing the selectivity of multidimensional range queries is a problem that arises in query optimization, data mining, and data warehousing. The query optimizer requires accurate estimations of the sizes of intermediate query results in the evaluation of different execution plans. Recent work also shows that top-\(k\) queries can be mapped to multidimensional queries [5,9]. Hence, selectivity estimation techniques can be used to optimize top-\(k\) queries as well.

In data mining, answering range queries is one of the simpler data exploration tasks. In this context, the user defines a specific region of the dataset that is worth exploring and asks queries to find the characteristics of this region (like the number of points in the interior of the region, the average value, or the sum of the values of attributes in the region). Consider for example a dataset that records readings of different environmental variables, such as types of pollution, at various space locations. In exploring this dataset, the user may be interested in answering range queries similar to: “find how many locations exist for which the values of given pollution variables are within a specified range.” The user may want to restrict the answers to a given geographical range, too. The size of such datasets makes exact answers intractable, and only an efficient approximation algorithm can make this data exploration task interactive.

In data warehousing, datasets are typically very large. Answering aggregate queries exactly can be computationally expensive. It is therefore very important to find approximate answers to aggregate queries quickly in order to allow the user to explore the data.

In this paper, we address the problem of estimating the selectivity of multidimensional range queries when the datasets have numerical attributes with real values. The range queries we consider are intersections of ranges, each range being defined on a single attribute. In the multidimensional attribute space, the queries are then hyperrectangles with faces parallel to the axes. Solving such a range query exactly involves counting how many points fall in the interior of the query. When the number of dimensions increases, recent results [39] show that the query time is linear to the size of the dataset.
We should emphasize that this problem is different from the traditional definition of selectivity estimation, where it has been generally assumed that each numerical attribute has a finite discrete domain. In many applications, however, the attribute domain is the real numbers. This is typically the case in spatial, temporal, and multimedia databases where the objects are represented as feature vectors (for example climatic data like humidity, wind speed, and precipitation).

Real domains have two important characteristics. First, the number of possible queries is infinite in the case of real domains but finite when considering a finite discrete domain. Of course, the number of possible distinct query answers is finite in both cases, since the dataset is finite. Second, with real domains it is unlikely that many attribute values will appear more than once in the database. Nevertheless, some techniques that have been developed to solve the discrete finite domain case are still applicable, if properly modified.

The main approach to solving the selectivity estimation problem has been to compute a nonparametric density estimator for the distribution function of the data. The methods suggested in the literature employ different techniques to find the density estimator for attributes with finite discrete domains. They include computing multidimensional histograms [29, 1, 20, 2], using the wavelet transformation [37, 23], SVD [29], the discrete cosine transform [22], or kernel estimators [3], and sampling [27, 21, 13].

Since the approximate solution to a query must be derived quickly, the description of the approximation function is kept in memory. Typically, an optimizer would maintain a separate approximation function for each of many datasets. Hence function descriptions cannot be very large. The success of the different methods depends on the simplicity of the function, the time it takes to find the function parameters, and the number of parameters stored for a given approximation quality.

Multidimensional density estimation techniques are typically generalizations of very successful one-dimensional density estimators. In general, in one dimension, estimators of small size (histograms, kernels, sampling) can be used to effectively approximate the data distribution. Indeed, one of the techniques used to solve the multidimensional problem is to assume that the attributes are independent, and therefore an estimator for multiple dimensions can be obtained by multiplying one-dimensional estimators.

Furthermore, we note that finding density estimators for combinations of attributes can be used to verify whether or not the independence assumption holds for a set of attributes. This is of independent importance for the query optimizer: many optimizers [32] compute query execution costs under the attribute independence assumption. If an optimizer can realize that this assumption does not hold for a set of attributes, more accurate statistics for this set should then be utilized.

1.1 Our contribution

In this paper, we give efficient techniques for finding density estimators for multidimensional datasets with real values. These techniques were originally introduced in [12] along with preliminary experimental results.

First, we describe GENHIST, an approach designed to find multidimensional histograms for datasets from real domains.

The basic feature of our technique is the overlapping of histogram buckets. Like other approaches, GENHIST uses more and smaller buckets to approximate the data distribution where the data density is higher and fewer and larger buckets where the density decreases. The difference is that these buckets are allowed to overlap, so that the data distribution at a given location is computed by considering all buckets that include this location.

Second, we show how to use multidimensional kernel density estimators to solve the multidimensional range query selectivity problem. Our solution generalizes in multiple dimensions the technique given in [3]. Kernel estimation is a generalization of sampling. Like sampling, finding a kernel estimator is efficient and can be performed in one pass. In addition, kernel estimators produce a smoother density approximation function, thereby producing a better approximation of the data density distribution.

Third, we present an extensive comparison between the proposed techniques (GENHIST and multidimensional kernel density estimators) and most of the existing approaches for estimating the selectivity of multidimensional range queries for real attributes (wavelet transform [37], multidimensional histogram MHIST [29], Min-Skew histograms [2], one-dimensional estimation techniques with the attribute independence assumption, and sampling [13]). We include the attribute independence assumption in our study as a baseline comparison. To the best of our knowledge this is the first work comparing a wide variety of selectivity estimators for multidimensional real-valued data.

The experimental results show that we can efficiently build selectivity estimators for multidimensional datasets with real attributes. Although the accuracy of all techniques drops rapidly as the dimensionality increases, the estimators are quite accurate up to ten dimensions. GENHIST is the most robust and accurate technique among the approaches that we have tested (MHIST, Min-Skew, kernels, sampling, and independence assumption) for space dimensionalities between three and ten. Among the other techniques, multidimensional kernel estimators are quite competitive with GENHIST in accuracy. An advantage of kernel estimators is that they can be computed in one dataset pass, just like sampling. However, they work better than sampling for the dimensionalities we tried. Therefore, multidimensional kernel estimators are the obvious choice when the selectivity estimator must be computed quickly.

In the next section (Sect. 2), we formally define the problem. In Sect. 3, we briefly describe the multidimensional histogram and wavelet decomposition approaches. GENHIST is introduced in Sect. 4, while Sect. 5 describes how to use kernel estimators for multidimensional data. Section 6 presents our experimental results, and Sect. 7 concludes the paper.

2 Problem description

Let \( R \) be a relation (dataset) with \( d \) attributes and \( n \) tuples. Let \( \mathcal{A} = \{A_1, A_2, \ldots, A_d\} \) be the set of these attributes. The domain of each attribute \( A_i \) is scaled to the real interval \([0, 1]\).

Assuming an ordering of the attributes, each tuple is a point in the \( d \)-dimensional space defined by the attributes. Let \( V_i \) be the set of values of \( A_i \) that are present in \( R \). Since the values