Using kernels to approximate multi-dimensional aggregate range queries over real attributes

Dimitrios Gunopulos [*]George Kollios [†]Vassilis J. Tsotras [‡]Carlotta Domeniconi [§]

Abstract

Approximating range queries: Finding approximate answers to multi-dimensional range queries over real valued attributes has significant applications in data exploration and database query optimization. In this proposal we consider the following problem: given a dataset of n points in the d-dimensional real space ($\Re^d$), and a query that specifies a range in each dimension, find a good approximation of the number of points in the dataset that satisfy the query, i.e. that fall within the specified d ranges.

Traditional Approaches: 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 multi-dimensional 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 which also affect the estimator.

[*] CS Dept., Univ. of California Riverside, Riverside, CA 92521, dg@cs.ucr.edu.

[†] CS Dept., Boston Univ., Boston, MA 02215, gkollios@cs.bu.edu.

[‡] CS Dept., Univ. of California Riverside, Riverside, CA 92521, tsotras@cs.ucr.edu.

[§] CS Dept., Univ. of California Riverside, Riverside, CA 92521, carlotta@cs.ucr.edu. Contact author.

Using Kernels: We show how to use multi-dimensional kernel density estimators to solve the multi-dimensional range query selectivity problem. 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 that better approximates the data density distribution, as shown experimentally. To further improve accuracy, we investigate the kernels technique to address the problem of setting the bandwidth parameters optimally.

Experiments and Results: We present an extensive comparison (using both synthetic and real data) between the new technique of multi-dimensional kernel density estimators and most of the existing techniques for estimating the selectivity of multi-dimensional range queries for real attributes (wavelet transform, multi-dimensional histogram MHIST, one-dimensional estimation techniques with the attribute independence assumption, and sampling). In our comparison we also include a recently presented histogram technique, GENHIST, that is designed to approximate the density of multi-dimensional datasets with real attributes. The technique finds buckets of variable size, and allows the buckets to overlap. The experimental results show that we can efficiently build selectivity estimators for multi-dimensional datasets with real attributes. Although the accuracy of all the techniques drops rapidly with the increase in dimensionality, the estimators are still accurate in 5 dimensions. Both multi-dimensional kernel estimators and GENHIST appear to be the most robust and accurate techniques among the ones we have tested. An advantage of kernel estimators is that they can be computed in one dataset pass (just like sampling). Moreover, they work better than sampling for the dimensionalities we have considered. Therefore, multi-dimensional kernel estimators are the best choice when the selectivity estimator must be computed fast.