Compressing spatio-temporal databases

December 23, 2013

1 Motivation

Suppose we want to study the effect of environmental variables, such as pollution and noise, on people's health. To collect data we use smart-phones instrumented with sensors that measure both the state of the environment (time, GPS-based location, temperture, noise, pollutants etc.) and the state of the person carrying them (heart-rate, coughing, oxygen in blood etc.)

Our goal is to identify the main dependencies between these variables. In order to reliably detect dependencies we need to bin/group/discretize/quantize time and space. Without such binning it is very difficult to relate different measurements to each other.

One of the challanges with binning is that spatial density is likely to be very uneven. The density of people in the mall on Christmas eve is much higher than the density of people on a remote mountain top at the same time.

Here we propose a data-driven binning method that adapts to the local density of measurements in time and location.

2 Data-driven partitions

Suppose we represent our measurements as triplets of the form (x, t, f) where x denotes location, t denotes time and f represents the vector of measurements taken at location x at time t.

Suppose we have collection of n measurements $(x_1, t_1, f_1), \ldots, (x_n, t_n, f_n)$. the simplest way of constructing a data-sensitive partition into m bins is to select m out of the n measurements at random. These m elements are called the "centroids". The measurement (x, t, f) is placed into the bin corresponding to the centroid whose location is closest to x in euclidean distance.

As the partition is randomized, it is often necessary to create several independent partitions and average the results over all of them.