Approximate Distance Methods for Clustering High-Dimensional Data

Lenore Cowen

Dept. of Math. Sc., Johns Hopkins University

Suppose $x_1, \ldots, x_n$ are a collection of n data points in $\Re^d$. The goal is to cluster this data into c classes. We are interested in this problem in the unsupervised case (where nothing is known a priori about the class labels, and we are essentially looking for shape and structure in a "point cloud") as well as in the supervised case, where some partial training data indicates correct class labels for some of the points, and the goal is to construct a rule that correctly classifies the remainder. For both these problems, there is a wealth of classical statistical literature for small d, but as d gets large, they are considered extremely hard– leading to what Hartigan/Gordon call "the curse of dimensionality".

There are many senses in which the curse of dimensionality is real, whether one takes a geometric approach based on orthogonal projections, or one focuses instead on capturing inter-point distances. A recent paper of Linial, London and Rabinovich, investigated algorithmic constructions for embedding points from $\Re^d$ into a lower dimensional Euclidean space, while approximately preserving all inter-point distances.[1] Their results are based on work of Bourgain, and Johnson and Lindenstrauss in the 1980s who showed that independent of the dimension of the original space, such points can be embedded in an $O(\log n)$ dimensional space, with at most a $O(\log n)$ distortion in any of the pairwise distances– and in some cases this is tight.

These results allows a tradeoff of approximation for dimension– in the case where $\log n$ is smaller than d, if we can tolerate the distortion, we reduce the curse of dimensionality. However, even if we can tolerate the distortion, for many applications, $O(\log n)$ dimensions is still too high to apply classical clustering and classification methods. Instead we propose a parametirized version of approximate distance dimension reduction, called ADC (for approximate distance clustering) where we introduce a family of maps based on distances to small random subsets, that preserve some, but not all pairwise distances at once. Note that if we are looking at a high-dimensional data set that clusters, we are mainly interested in preserving the inter-class distances, if the intra-class distances collapse, this only increases the clustering in the low-dimensional space.

The methods involve both the combinatorics of these random subsets, and resulting rank sum statistics, and a theoretical computer science approach to randomized algorithms.

(Joint work with C.E. Priebe)

[1]Many of their results generalize to other distance metrics.