An Improved Clustering Algorithm

Talya Meltzer and Jonathan S. Yedidia
Disney Research Boston,
222 Third St., Cambridge, MA 02142 USA
{talya.meltzer, yedidia}@disneyresearch.com
http://www.disneyresearch.com

We present a new clustering algorithm that improves upon the very well-known and widely-used Lloyds algorithm for K-means clustering. The clustering problem is to partition a set of data points into clusters such that the points in each cluster are as similar as possible to each other. In the standard K-means formulation of the clustering problem, all data-points in a cluster are represented by a single centroid point, and the goal is to minimize the distortion, as measured by the sum of squared distances of each data point to its associated centroid, where the number of clusters (K) is chosen ahead of time.

The K-means clustering problem is NP-hard, and the most popular approach to dealing with it is the heuristic Lloyds algorithm, which quickly finds a locally optimal solution. Lloyds algorithm aims to minimize the distortion by iteratively repeating two steps: first, given a partition of the data points, optimize the centroid locations by placing them at the mean position of the data points in the cluster; and secondly, given a set of centroid positions, optimize over the partitions by associating each data point with the nearest centroid. The algorithm terminates when there is no change in the partitioning. Despite its enormous popularity, the Lloyds algorithm is highly non-optimal, and will typically find solutions that are bad local minima in the distortion.

In this work, we improve upon the performance of the standard Lloyds algorithm by making a simple change to its second step. When choosing which centroid to associate with each data point, we choose, with a probability of approximately 40%, the second-nearest centroid if its distance to the data point is no more than $(1+T)$ times as that of the nearest centroid. The parameter T is normally initialized to some large value, and decreased geometrically towards zero with each iteration. This algorithm, which we call the *T-Lloyds algorithm,* terminates when no change is possible in the partitioning, which always happens when T becomes sufficiently small. It reduces to the standard Lloyds algorithm when T is initialized to zero. The T-Lloyds algorithm superficially resembles simulated annealing, but differs in the fact that the configurations completely freeze when T becomes small enough.

We have experimented with the T-Lloyds algorithm on different sets of problems where the exact optimal solution to the clustering problem is known, so that

we can evaluate the significance of the improvement. Although the T-Lloyds algorithm runs for more iterations than standard Lloyds (typically around 20 times as many), the improvement it achieves is very significant. In particular, in practice, because of its non-optimality, the standard Lloyds algorithm is often repeated many times, and the best solution is chosen. We show that even running T-Lloyds a smaller number of times, to account for the greater number of iterations it uses, will still give a much better result than using the Lloyds algorithm.

Topic: learning algorithm
Preference: poster