Probabilistic distance clustering

Adi Ben-Israel and Cem Iyigun

The State University of New Jersey, USA

Abstract

Distance (center based) clustering methods iterate between determining the clusters (or centers) and (re)assigning the data points to these clusters. The work reported here uses soft assignments, with probabilities assumed to depend on distances, for example, $p(x|k)d(x|k) = D(x)$, for all data points x, where $p(x|k)$ is the probability that x belongs to the k-th cluster, $d(x|k)$ its distance from that cluster, and $D(x)$ independent of k. The function $D(x)$ measures the classifiability of the data in question (it is, up to a constant, the harmonic mean of the distances $d(x|k)$). The probabilities $p(x|k)$ are not needed computationally (all results, in particular the centers updates, are stated in terms of the distances).

Results:

1. A new distance clustering method [2], [3].
2. A viable alternative to the EM method for demixing distributions.
3. The "right" number of clusters for a given data set.
4. Contour approximation, and compact representation, of data, [1], [4].
5. A modern optimization framework [5].
6. A generalization of theWeiszfeld method for solving the Fermat-Weber location problem with several centers
7. Semi supervised clustering, using prior information (labels). In particular, why any classification method works well on some ("good") data sets, and how to decide if a data set is "good".

References

M. Arav (2008). Contour approximation of data and the harmonic mean. *Math. Ineq. & Appl.*. To appear.

A. Ben-Israel and C. Iyigun (2007). Probabilistic distance clustering. *J. Classification*. To appear.
URL: http://benisrael.net/J-CLASSIFICATION-07.pdf

C. Iyigun and A. Ben-Israel (2007). Probabilistic distance clustering adjusted for cluster size. *Probability in Engineering and Informational Sciences*. To appear.
URL: http://benisrael.net/PEIS-07.pdf

C. Iyigun and A. Ben-Israel (2007). Contour approximation of data: A duality theory.
URL: http://benisrael.net/DUAL-12-20-07.pdf

M. Teboulle (2007). A unified continuous optimization framework for center-based clustering methods. *J. of Machine Learning*, *8*, 65-102.