Sean Kelley

Machine Learning – Prof. Brodley

Constrained K-means Clustering with Background Knowledge

K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl

This paper focuses on the shortcomings of the popular k-means algorithm and proposes a solution to fix a particular area that is lacking. Despite its popularity and frequency of use, k-means has no formal mechanism for incorporating potentially valuable background information, namely information on partial clusterings (constraints). By letting users of the algorithm manually or programmatically flag data points as definitely (not) belonging to the same cluster, the algorithm’s accuracy is dramatically increased.

A major advantage of the paper is that it proposes an easy-to-use, easy-to-implement solution to a common problem. There is no reason to believe that this algorithm improvement would not be immediately ready to be applied to real-world scenarios; indeed, they describe one such scenario in the paper itself. There appears to be no performance penalty for not supplying the requested constraints, and any such information is guaranteed to be preserved: in any domain where k-means is used, this should be likely used as well.

My primary criticism while reading the example section of the paper was that k-means is an algorithm unsuitable for accurately solving the presented problem, but the authors have already responded with a suggestion: that such unsuitable algorithms become less of a liability when combined with the power of constraints. This may be an area worth exploring, namely, what the payoffs of using a generalized clustering algorithm with constraints is versus exploring and discovering a more suitable clustering algorithm and integrating constraints into it.

The algorithm’s blind trust of the supplied constraints may be a liability, however. If the constraints are misused they have the potential to break the algorithm. In one dataset, the algorithm erroneously assigned a data point the wrong constraints and performed worse for it. Admittedly it still dramatically outperformed regular k-means, though a systematic mistake on the part of the background knowledge supplier (human or machine) could conceivably be a detriment rather than a benefit. Again, more testing is needed to weight the benefits of this approach when your a *priori* knowledge supplier is noisy.