A *Variations* Tool for Clustering

Rich Caruana, Mohamed Elhawary, Nam Nguyen, Casey Smith
Department of Computer Science, Cornell University, Ithaca, NY 14853
caruana@cs.cornell.edu www.cs.cornell.edu/~caruana

Photo editors such as Photoshop have a tool that presents users with variations of their picture with different color balances, color saturations, and brightnesses. Instead of having to know what tools, filters, and parameters to use to improve the picture (which requires substantial expertise and trial and error), the user only has to select the variation that looks best. The selected variation then becomes the new center, and variations of it are presented, allowing the user to quickly zero in on the desired rendition.

We have built a "variations" tool for clustering. Most clustering users are not clustering experts and just want to use clustering as a tool in their work. They don't want to modify clustering algorithms, try different clustering methods, or spend too much time tweaking the similarity metric to obtain a clustering that is useful for their problem. They want clustering to be easier, *and* to return better clusterings. The variations tool we developed allows users to select (and optionally refine) a clustering that is best for their purposes from a variety of different, automatically generated clusterings. The automatically generated clusterings are hierarchically organized so that users can quickly zero in on good clusterings without having to look at many alternatives.

The clustering variations tool is based on a stochastic clustering method based on iterated k-means and spectral clustering. Clustering variations are found through a combination of PCA and random projections. Clustering variations are hierarchically organized by clustering the clusterings at a meta-level using a distance measure defined over entire clusterings [5], and the number of clusterings is further reduced and their quality improved by merging similar clusterings with consensus clustering [1, 3].

Experiments with this new clustering tool on six data sets has produced a few interesting results:

- The most useful clusterings often are not very compact, and algorithms that find compact clusterings often do not find useful clusterings. The variations tool often finds *much* better clusterings.
- The clustering of the data that is best for one purpose often is *very* different from the clustering of that same data that is best for a different purpose. There is no one *correct* clustering.
- It is more important to get the distance metric right than to get the clustering algorithm right.
- k-means with random restarts does not generate a very diverse set of clusterings on many data sets.
- To get good results we had to assume that feature importance is distributed as a Zipf distribution [2].
- If you can't afford to generate many different clusterings and have to run just one clustering algorithm, the spectral clustering method developed by Ng, Jordan, and Weiss [4] seemed to yield the best results on our data.

References

[1] C. Boulis and M. Ostendorf. (2004) Combining multiple clustering systems. *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases.*
[2] S. Cohen and G. Dror and E. Ruppin. (2005) A Feature Selection Method Based on the Shapley Value. *Proceedings of the International Joint Conference on Artificial Intelligence.*
[3] V. Filkov and S. Skiena. (2003) Integrating microarray data by concensus clustering. *Proceedings of the International Conference on Tools with Artificial Intelligence.*
[4] Andrew Y. Ng, Michael Jordan, and Yair Weiss. (2002) On Spectral Clustering: Analysis and an algorithm. *NIPS.*
[5] L. Hubert and P. Arabie. (1985) Comparing partitions *Classification* Vol. 2.

Topic: Clustering
Preference: Oral