Theoretical Foundations of Clustering

Saturday December 10, 2005

Shai Ben-David, *University of Waterloo, Canada*
Ulrike von Luxburg, *Fraunhofer IPSI, Germany*
John Shawe-Taylor, *University of Southampton, UK*
Naftali Tishby, *Hebrew University, Israel*

http://www.ipsi.fraunhofer.de/∼ule/clustering_workshop_nips05/workshop_description.html

Background: Clustering is one of the most widely used techniques for exploratory data analysis. Despite the large number of algorithms and applications, the theoretical foundations of clustering seem to be distressingly meager, covering only some sub-domains and failing to address some of the most basic general aspects of the area. We wish to initiate a concerted discussion, in order to move towards a consolidation of the theoretical basis for - at least some of the aspects of - clustering. One prospective benefit of building a theoretical framework for clustering may come from enabling the transfer of tools developed in other related domains, such as machine learning and information theory, where the usefulness of having a general mathematical framework have been impressively demonstrated.

Questions we wish to address:

1. What is clustering? How can it be defined?
   - Is the main purpose of clustering to discover new features in the data? To simplify our data by building groups, thus getting rid of unimportant information?
   - Is clustering just data compression? Is clustering just estimating modes of a density?
   - Is clustering related to human perception?
   - Can one come up with a meaningful taxonomy of clustering tasks?
2. How should prior knowledge be encoded? As a pair-wise similarity/distance function over domain points? As a set of relevant features? Should data be embedded in a richer structure?
3. Is there a principled way to measure the quality of a clustering on particular data set?
   - Can *every* clustering task be expressed as an optimization of some explicit, readily computable, objective cost function?
   - Can stability be considered a first principle for meaningful clustering?
4. Is there a principled way to measure the quality of a clustering algorithm? What are sufficient conditions for reasonable clustering? Are there obvious necessary conditions? What type of performance guarantees can one hope to provide?
5. How should the similarity between different clusterings be measured?
6. Can one distinguish clusterable data from structureless data?
7. What are the tools we should try to import from other relevant areas of research?

Theoretical Foundations of Clustering

Saturday December 10, 2005

Organizers: Shai Ben-David& Ulrike von Luxburg & John Shawe-Taylor & Naftali Tishby

Morning session: 7:30am–10:30am

7:30am **Introduction and Goals of the Workshop**, *Ulrike von Luxburg*

What is clustering?

7:45am **Attempts to formalize clustering**, *Shai Ben-David*

8:15am **On the futility of attempts to formalize clustering within the conventional mathematical framework**, *Lev Goldfarb*

8:25am **Informational and computational limits of clustering**,
Nati Srebro, Sam Roweis, and Gregory Shakhnarovich

8:35am *discussion*

8:45am *coffee break*

Stability and Evaluation of clustering

8:55am ***Invited talk:*** **Confidence and stability in comparing clusterings**, *Marina Meila*

9:25am **Cascade evaluation**,
Laurent Candillier, Isabelle Tellier, Fabien Torre, and Olivier Bousquet

9:35 ***Invited talk:*** **Data Clustering and the Stability Method**, *Joachim M Buhmann*

10:05 **Stability of clustering method**, *Sasha Rakhlin*

10:15 *discussion*

Afternoon session: 3:30pm–6:30pm

The clustering input structure

3:30pm **Feature space generalization - the missing dimension of learning?**, *Tali Tishby*

4:00pm ***Invited talk:*** **Learning with similarity functions**, *Avrim Blum*

4:30 **Learning clusterwise similarity with first order formulas**,
Aron Culotta and Andrew McCallum

4:40pm *coffee break*

Information theoretic and other principled approaches to clustering

4:50pm ***Invited talk:*** **Clustering from a rate-distortion perspective**, *Joydeep Ghosh*

5:20pm **Information based clustering - a principled approach for cluster analysis**, *Noam Slonim, Gurinder Singh Atwal, Gasper Tkacik, William Bialek*

5:30pm **An MDL Framework for Data Clustering**, *Petri Myllymaki*

5:40pm **Density traversal clustering**, *Amos J Storkey and Tom G Griffiths*

5:50pm **Clustering and Staircases**, *K. Pelckmans, J.A.K. Suykens, B. De Moor*

6:00pm *Concluding discussion*