NIPS 2005 Workshop *Whistler BC, Saturday, December 10*

Theoretical Foundations of Clustering

For the workshop updated site see

www.ipsi.fraunhofer.de/%7Eule/clustering_workshop_nips05/clustering_workshop_nips05.htm

Organizers:

Shai Ben-David, Ulrike von Luxburg, John Shawe-Taylor and Naftali Tishby

Background:

Clustering is one of the most widely used techniques for exploratory data analysis. Across all disciplines, from social sciences over biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. In the past five decades, a wide variety of clustering algorithms have been developed and applied to a wide range of practical problems.

Despite this large number of algorithms and applications, the goal of clustering and its proper interpretation remains fuzzy and vague. There are in fact many different problems that are clustered together under this single term, from quantization with low distortion for compression, through various techniques for graph partitioning whose goals are not fully specified, to methods for revealing hidden structure and unobserved features in complex data. We are clearly not talking about a single well defined problem.

Moreover, 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. There is not even an agreement among the researchers on the correct questions to pose, let alone which tools and analysis techniques should be used to answer those questions.
In our opinion there is an urgent need to initiate a concerted discussion on these issues, 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.

We have the impression that recently many researchers have become aware of this need and agree on the importance of these issues.
In part, the idea of having this workshop has its origin in a workshop on the (ab)use of bounds that some of us have organized at NIPS'04, where it turned out that a significant portion of the talks and discussions revolved around unresolved fundamental issues of clustering.

Some specific workshop topics:

We wish the workshop to address questions like:

- What is clustering? How can it be defined and how can we sort the different types of clustering and their goals?
  In particular:
  - Is the main purpose to use the partition to discover new features in the data?
  - Or the other way around, is the main purpose 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?
  - Can we formulate the intuitive notion of "revealing hidden structure and properties"?

- 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 some richer structure (Hilbert space, topology) ?
- 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 associated objective cost function?
  - Can stability be considered a first principle for meaningful clustering?

- Is there a principled way to measure the quality of a clustering algorithm?
  - Necessary conditions
  - Can we come up with sufficient conditions for reasonable clustering?
  - Stability conditions
  - Richness conditions
  - What type of performance guarantees can one hope to provide?

- What are principled and meaningful ways of measuring the similarity (or degree of agreement) between different clusterings?

- Can one distinguish “clusterable” data from “structureless” data?

- What are the tools we should try to import from other areas such as classification prediction, density estimation, data compression, computational geometry, other relevant areas?

Structure of the workshop:

We plan a 1-day workshop. We will start by an introductory tutorial raising and defining the questions mentioned above. Then we will have about 4 sessions, each dedicated to one of the topics mentioned above.
The idea is that in each session we have (very few) short talks representing different views on the current question.
The core part of each session will be a discussion of the different arguments presented in the talks.

Finally, at the end of the workshop the organizers will summarize the different issues that have been raised, highlighting points on which consensus has been reached and listing the different views on topics which are still controversial.

Organization:

The workshop organizers will serve as a program committee that will invite and evaluate proposals for short talks from authors with different viewpoints.
We shall evaluate proposed talks on the merits of both their relevance and significance and their contribution to the well-roundedness of the workshop as a whole.

Workshop goals:

We hope that the workshop will serve as a starting point for focused research on the foundations of clustering, stimulating collaborations and contributing to the awareness of the participants and the NIPS community at large about the different aspects of and approaches to the development of a general clustering theory.

Organizers' web pages:

Shai Ben-David (www.cs.uwaterloo.ca/~shai)
Ulrike von Luxburg (www.ipsi.fraunhofer.de/mine/en/people/luxburg)
Naftali Tishby (www.cs.huji.ac.il/~tishby)
John Shawe-Taylor (www.ecs.soton.ac.uk/people/jst)