Fall term, 2006

CS886 - Theoretical Foundations of Clustering

Shai Ben-David,

http://www.cs.uwaterloo.ca/~shai

Schedule and Location: We'll meet once a week, **Thursday 2:30-5:30 in DC3313**

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.

Recently many researchers have become aware of this need and agree on the importance of these issues. There have been a number of recent workshops devoted to this topic and there are quite a few recent research papers addressing a variety of basic questions about clustering from a theoretical perspective.

Some example issues that we may address:

- 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 course:

The course will consist mostly of paper presentations followed by discussions. I shall start by giving several introductory lectures but the major part of the course will consist of student presentations. The presentations will be based of research papers (I am not

aware of any text book that covers a large enough portion of the topics I wish to discuss). The papers to be discussed will be selected by the participating students from a list of paper that I shall introduce, as well as papers suggested by the speakers (subject to my approval, of course).

Prerequisites:

There is no formal prerequisite but I expect participants to have some interest in clustering as well as a positive attitude towards theory and abstract mathematics. The most relevant mathematical areas are Probability and Statistics, Linear Algebra and Combinatorics.

Requirements and workload:

1) I expect students to read the relevant papers towards each lecture (to the point of being able to write a short summary of the paper – technical understanding of proofs is required only from the student presenting the paper).
   All students should submit short summaries of the papers to be discussed.
2) Each student is expected to thoroughly read and present at least one research paper.
3) Each student will submit a small project – either a detailed critical review of a published paper or an outline of an original research.

Grade structure:

1) Summaries of papers – 10%
2) Class participation – 10%
3) Student lecture- 60%
4) Project – 20%