Correlation Clustering: from Theory to Practice

Francesco Bonchi
Yahoo Labs
Barcelona, Spain
bonchi@yahoo-inc.com

David García-Soriano
Yahoo Labs
Barcelona, Spain
davidgs@yahoo-inc.com

Edo Liberty
Yahoo Labs
New York City, USA
edo@yahoo-inc.com

ABSTRACT

Correlation clustering is arguably the most natural formulation of clustering. Given a set of objects and a pairwise similarity measure between them, the goal is to cluster the objects so that, to the best possible extent, similar objects are put in the same cluster and dissimilar objects are put in different clusters. As it just needs a definition of similarity, its broad generality makes it applicable to a wide range of problems in different contexts, and in particular makes it naturally suitable to clustering structured objects for which feature vectors can be difficult to obtain.

Despite its simplicity, generality and wide applicability, correlation clustering has so far received much more attention from the algorithmic theory community than from the data mining community. The goal of this tutorial is to show how correlation clustering can be a powerful addition to the toolkit of the data mining researcher and practitioner, and to encourage discussions and further research in the area.

In the tutorial we will survey the problem and its most common variants, with an emphasis on the algorithmic techniques and key ideas developed to derive efficient solutions. We will motivate the problems and discuss real-world applications, the scalability issues that may arise, and the existing approaches to handle them.

Target audience and prerequisites

The tutorial is aimed at researchers interested in the theory and applications of clustering. No special knowledge will be assumed other than familiarity with algorithmic techniques from a standard computer science background.

Outline

The tutorial is structured in three main technical parts, plus a concluding part where we will discuss future research agenda. All three technical parts will contain both theory and real-world applications.

1. **Introduction and fundamental results:** Motivating examples; problem formulation; hardness; maximization and minimization versions; relationship with agnostic learning; approximation algorithms; applications.

2. **Correlation clustering variants:** Clustering aggregation/consensus clustering; bipartite correlation clustering; overlapping correlation clustering; chromatic correlation clustering; online correlation clustering; existence of a ground truth clustering; planted partition models; random noise; connections to other problems.

3. **Scalability for real-world instances:** Empirical approaches to overcome the all-pairs barrier; sublinear query complexity via relative regret approximations; neighborhood oracles; local correlation clustering; examples of real-world applications on Big Data.

4. **Challenges and directions for future research**

Instructors

Francesco Bonchi is leading the "Web Mining" research group at Yahoo Labs in Barcelona, Spain. He is member of the ECML PKDD Steering Committee, member of the Editorial Board of ACM Transactions on Intelligent Systems and Technology (TIST), and Associate Editor of IEEE Transactions on Knowledge and Data Engineering (TKDE). He has been program co-chair of ECML PKDD 2010, PinKDD 2007 and 2008, PADM 2006, and KDID 2005. More information can be found at
`http://www.francescobonchi.com/`

David García-Soriano is a postdoctoral researcher at Yahoo Labs in Barcelona. He received his PhD from the University of Amsterdam under the supervision of Harry Buhrman. His research interests include sublinear-time algorithms, learning, approximation algorithms, and large-scale problems in data mining and machine learning. More information can be found at
`https://sites.google.com/site/elhipercubo/`

Edo Liberty is leading the "Scalable Machine Learning" research group at Yahoo Labs in New York. He received his PhD in Computer Science from Yale University, under the supervision of Steven Zucker. After that, he joined the Program in Applied Mathematics at Yale as a Post-Doctoral fellow. In 2009 he joined Yahoo Labs. He received the best paper award at SODA 2011 and KDD 2013. More information can be found at `http://www.cs.yale.edu/homes/el327/`

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2630808.