How Far Are We From Having a Satisfactory Theory of Clustering?

Shai Ben-David[1]

1 **Universitys of Waterloo, Waterloo, Ontario, Canada**
shai@cs.uwaterloo.edu

Abstract

This is an overview of the invited talk delivered at the 41st International Symposium on Mathematical Foundations of Computer Science (MFCS-2016).

1998 ACM Subject Classification I.5.3 Clustering

Keywords and phrases clustering, theory, algorithm tuning, computational complexity

Digital Object Identifier 10.4230/LIPIcs.MFCS.2016.1

Category Invited Talk

1 Overview of the Talk

Unsupervised learning, utilizing the huge amounts of raw data available, is widely recognized as one of the most important challenges facing machine learning nowadays. For supervised tasks, machine learning theory has been successful in several respects; providing significant understanding of machine learning tasks (in terms of the informational and computational resources they require and in providing algorithmic tools to address them), insights about the pros and cons of alternative machine learning paradigms and their parameter settings, and initiating the development of new algorithmic approaches. However, no such successes had been achieved so far for the unsupervised ML domain.

My talk will focus on clustering, arguably the most fundamental unsupervised data processing task. I will discuss two aspects in which theory could play a significant role in guiding the use of clustering tools. The first is model selection - how should a user pick an appropriate clustering tool for a given clustering problem, and how should the parameters of such an algorithmic tool be tuned? In contrast with other common computational tasks, in clustering, different algorithms often yield drastically different outcomes. Therefore, the choice of a clustering algorithm may play a crucial role in the usefulness of an output clustering solution. Just the same, currently there exist no methodical guidance for clustering tool selection for a given clustering task. I will describe some recent proposals aiming to address this crucial lacuna.

The second aspect of clustering that I will address is the computational complexity of computing a cost minimizing clustering (given some clustering objective function). Once a clustering model (or objective) has been picked, the task becomes an optimization problem. While most of the clustering objective optimization problems are computationally infeasible, they are being carried out routinely in practice. This theory-practice gap has attracted significant research attention recently. I will describe some of the theoretical attempts to address this gap and discuss how close do they bring us to a satisfactory understanding of the computational resources needed for achieving good clustering solutions.