Clustering Quality Measures

Margareta Ackerman and Shai Ben-David
D.R.C. School of Computer Science
University of Waterloo
{mackerma, shai}@cs.uwaterloo.ca
www.cs.uwaterloo.cs/∼ shai

Abstract

Aiming towards the development of a general clustering theory, addressing issues that are common to the different clustering paradigms, we wish to initiate a systematic study of measures for the *quality of a given data clustering*. A clustering quality measure is a function that, given a data set and its partition into clusters, returns a non-negative real number representing the quality of that clustering. We analyze what clustering quality measures should look like by introducing a set of requirements ('axioms') of clustering quality measures. We propose quality measures for wide families of common clustering approaches, like loss-based clustering, center-based clustering, and linkage-based clustering. We show that our proposed measures satisfy the axioms and analyze their computational complexity.

1 Introduction

In his highly influential paper, [1], Kleinberg advocates the development of a theory of clustering that will be "independent of any particular algorithm, objective function, or generative data model". As a step in that direction, Kleinberg sets up a set of "axioms" aimed to define what a clustering function is. Kleinberg suggests three axioms, each sounding plausible, and shows that these seemingly natural axioms lead to a contradiction - there exist no function that satisfies all three requirements. As noted in the last section of [1], that "impossibility theorem" applies only to a very specific set of axioms. Small changes to any of these axioms suffices to turn them into a consistent set of requirements that are met by many common clustering paradigms. Just the same, Kleinberg's result is often interpreted as stating the impossibility of defining what clustering is, or even of developing a general theory of clustering.

We take up a similar line of research - aiming to develop a high level theory of clustering, investigating an axiomatic approach. However, rather than attempting to define what a *clustering function* is, and demonstrating a failed attempt, we turn our attention to the closely related issue of determining the *quality of a given data clustering* and come up with a consistent formalization of that notion.

The aim of clustering is to uncover meaningful groups in data. However, not any arbitrary partitioning of a given data set reflects such a structure. Upon obtaining a clustering, usually via some algorithm, a user needs to determine whether this clustering is sufficiently meaningful to rely upon for further data mining analysis or practical applications. That is, a user needs to be able to judge how good is a specific clustering. Clustering quality measures can also be used to compare different clusterings over the same data set. Different clustering algorithms aim to optimize different (potentially implicit) objective functions and are likely to output different clusterings of the same data set. Since it is often ambiguous which loss function, if any, is appropriate for clustering the data set at hand, a user may choose to try a number of different algorithms. Furthermore, most clustering algorithms, rely on the user to tune clustering parameters, like the number of clusters or a pruning rule. Therefore, a user needs a way to compare the quality of clusterings obtained by choosing different values of these parameters even when applying a fixed clustering paradigm. Clustering quality measures should provide a principled method for comparing clusterings and for evaluating their significance. We formalize the process of clustering quality evaluation by studying clustering quality measures.

When posed with the problem of finding a clustering quality measure, a first attempt may be to invoke the loss (or objective) function used by the clustering al-

gorithm, such as k-means or k-median, as a clustering quality measure. However, such measures have some major shortcomings for the purpose at hand. First, they are usually not scale-invariant. Given any non-trivial data partitioning (where at least one cluster has at least two points), any k-means or k-median loss can be obtained, for that fixed partitioning, by uniformly scaling the pairwise distances between points in the underlying data set. Consequently, by knowing that the k-means cost of some clustering is, say 7.3, one gains no insight about the quality, or "meaningfulness", of that clustering. Furthermore, clustering objective functions are usually sensitive to the number of clusters, and thus cannot be readily applied for choosing among clusterings with a different number of clusters.

Our goal in this paper is to formulate a theoretical basis for clustering quality evaluation. To our best knowledge the concept of a clustering quality measure has not been previously formalized. We address the question of what a measure of clustering quality should look like. We propose a set of requirements ('axioms') of clustering quality measures. We introduce several such quality measures for several common clustering paradigms, including loss-based clustering, center-based clustering, and linkage-based clustering. We compare these different notions of clustering quality and show that they all satisfy our axioms.

References

[1] Jon Kleinberg. "An Impossibility Theorem for Clustering." Advances in Neural Information Processing Systems (NIPS) 15, 2002.