Measures of Clustering Quality: A Working Set of Axioms for Clustering

Margareta Ackerman and Shai Ben-David
D.R.C. School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
{mackerma, shai}@uwaterloo.ca

Abstract

Aiming towards the development of a general clustering theory, 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 how 'strong' or 'conclusive' the clustering is. We propose using this notion as a basis for developing a formal theory of clustering. We analyze what clustering-quality measures should look like by introducing a set of requirements ('axioms') of clustering-quality measures. As opposed to previous work focusing on clustering functions as the object to be axiomatized, we show that principles like those formulated in Kleinberg's axioms ([2]) can be readily expressed in our framework without leading to inconsistency.

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. In addition, we show that using our measures, the clustering quality of a clustering can be computed in low polynomial time.

1 Introduction

In his highly influential paper, [2], 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 exists no function that satisfies all three requirements. As noted in the last section of [2], this "impossibility theorem" applies only to a very specific set of axioms. Small changes to any of these axioms suffice 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 disagree with this view.

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 evaluating the *quality of a given data clustering* and come up with a consistent formalization of that notion.

As it turns out, the clustering-quality framework is richer and more flexible than that of clustering functions. It allows the postulation of axioms that capture the features that Kleinberg axioms aim to express, while maintaining consistency of the set of axioms.

A *clustering-quality measure* is a function that maps pairs of the form (*dataset*, *clustering*) to some ordered set (say, the set of non-negative real numbers), so that these values reflect how 'good' or 'cogent' that clustering is.

The need to measure the quality of a given data clustering arises naturally in many clustering issues. 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. Clustering-quality measures judge how good is a specific clustering.

Clustering-quality measures can also be used to help in clustering model selection by comparing 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 objective function, if any, is appropriate for clustering the data set at hand, a user may apply a clustering-quality measure to choose between the outcomes of different algorithms, or between the outcomes of the same algorithm under different parameter settings. Clustering-quality measures should provide a principled method for comparing clusterings and for evaluating their significance.

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 algorithm, 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 loss of some clustering is, say 7.3, one gains no insight about the quality, or "meaningfulness," of that clustering. Second, the value of a specific loss function cannot be meaningfully used to compare the quality of clusterings obtained by different algorithms. If such quality measure is used, then an algorithm that aims to minimize that loss function has an advantage that is not necessary reflected in the quality of the clusterings that it produces. For example, the k-means loss function cannot be used to meaningfully compare a clustering obtained using a k-means heuristic with a clustering obtained via an algorithm aimed at optimizing a different objective function.

We formulate a theoretical basis for clustering-quality evaluations. To the best of our knowledge, there is no previously published formalization of such a notion. We address the question of what a measure of clustering quality should look like and propose a set of requirements ('axioms') of clustering-quality measures. To demonstrate the relevance and consistency of these axioms, we introduce concrete quality-measures for several common clustering paradigms, including loss-based clustering, center-based clustering, and linkage-based clustering. These notions all satisfy our axioms, and, given a data clustering, their value on that clustering can be computed in low polynomial time.

We begin by presenting Kleinberg's axioms for clustering functions. Next, we show that principles like those formulated in these axioms can be readily expressed as a satisfiable set of properties in the framework of quality measures. However, some of these properties are too strong to be used as axioms. Using these properties as a starting point, we derive our axioms of clustering-quality measures. Next, we present our clustering-quality measures, starting with measures for loss-based clustering, where we show a number of ways to normalize a loss function to obtain meaningful clustering-quality measures. In addition, we discuss how spectral clustering relates to loss-based quality measures, and present a quality measure based on spectral clustering theory. We then present quality measures for center-based and linkage-based clustering. Finally, we show that using each of our quality measures, the quality of a clustering can be computed in low polynomial time.

1.1 Other possible primitives for clustering formal analysis

There are various basic notions that one can use as the primitive notion for a discussion of the fundamental properties of clustering. As mentioned above, Kleinberg [2] has chosen "clustering functions" as that primitive for his analysis. Another basic notion that has been discussed in the literature is "clusterability" (for example, [5] and [6]). A notion of clusterability determines how much 'clustered structure' there is in a data. The notion of clustering-quality that we discuss is closely related to these other notions. A clustering-quality measure can readily induce a clustering *function* by assigning to any input data set the highest quality clustering of that set (whenever the data set is finite, then so is the set of all its clusterings. Thus a clustering with maximal clustering quality exists). Similarly, having a clustering-quality measure, one can define the *clusterability* of a data set as the maximal value of the clustering-quality over all clusterings of that set.

On the other hand, notions of clusterability cannot be used directly as measures of clustering quality, since clusterability is a property of a data set while measures of clustering-quality apply to specific clusterings.

2 Formal Framework and Notation

Let $X = \{1, 2, \ldots, n\}$ be a data set. A function $d : X \times X \to \mathbf{R}$ is a *distance function* if $d(x_i, x_i) \geq 0$ for all $x_i \in X$, for any $x_i, x_j \in X$, $d(x_i, x_j) > 0$ if and only if $x_i \neq x_j$, and $d(x_i, x_j) = d(x_j, x_i)$. Note that we do not require the triangle inequality.

A *k-clustering* $C = \{C_1, C_2, \ldots, C_k\}$ of data set X is a k-partition of X, that is, $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^k C_i = X$. A *clustering* of X is a k-clustering of X for some $k \geq 1$. A clustering of a data set on n

elements is *trivial* if it consists of 1 or n non-empty clusters.

For $x, y \in X$ and clustering C of X, we write $x \sim_C y$ whenever x and y are in the same cluster of clustering C and $x \not\sim_C y$, otherwise.

A *clustering-quality measure* is a function that is given a clustering C over (X, d) (where d is a distance function over X) and returns a non-negative real number. In this work we discuss what requirements are necessary to make such function a meaningful clustering-quality measure.

3 Axiomatizing clustering functions vs clustering quality measures

Previous attempts to provide an axiomatic basis for clustering were mostly carried out using clustering functions as the basic object of discussion [2], [1] or addressing clustering objective functions as the central object of discussion [4]. Since Kleinberg's framework is the closest to what we are doing here, we begin by describing his proposed axioms (for clustering functions).

3.1 Axioms for Clustering Functions

A *clustering function* for some domain set X is a function f that takes a distance function d over X, and outputs a partition of X. Kleinberg proposes three axioms of clustering functions, which he shows to be inconsistent [2]. We discuss Kleinberg's axioms, and how they relate to our work on clustering-quality measures.

Function Scale Invariance: Scale invariance requires that the output of a clustering function be unaffected by uniform scaling of the input.

A function f is scale-invariant *if for every distance function d and positive λ, $f(d) = f(\lambda d)$ (where λd is defined by setting, for every pair of domain points x, y, $\lambda d(x, y) = \lambda \cdot d(x, y)$).*

Function Consistency: Consistency requires that if within-cluster distances are decreased, and between-cluster distances are increased, then the output of a clustering function does not change. Formally,

- A distance function d' is a *C-consistent variant* of d, if $d'(x, y) \leq d(x, y)$ for all $x \sim_C y$, and $d'(x, y) \geq d(x, y)$ for all $x \not\sim_C y$.
- *A function f is* consistent *if $f(d) = f(d')$ whenever d' is an $f(d)$-consistent variant of d.*

Function Richness: Richness requires that by modifying the distance function, any partition of the underlying data set can be obtained.

A function f is rich *if for each clustering C of X, there exists a distance function d over X so that $f(d) = C$.*

One can readily transfer the Kleinberg axioms to the context of clustering-quality measures. Namely, given a quality measure, m, define a clustering function F_m by

$$F_m(X, d) = Argmax_C\{m(C, X, d)\}$$

That is, F_m selects the clustering of maximum quality. Now, given any requirement on clustering *functions*, one can transform it into a requirement on *clustering-quality measures* by stating that the induced function, F_m, satisfies this requirement. For instance,

Definition 1 *A quality measure m is* K-consistent *if for any $F_m(X, d)$ consistent variant d' of distance function d, $F_m(X, d) = F_m(X, d')$.*

We shall denote the quality-function requirements resulting from such direct translation of Kelinberg's axioms by *K-consistency*, *K-richness* and *K-scale-invariance* (where the "K" stands for "Kleinberg").

Clearly, if we start with an unsatisfiable set of axioms for clustering functions, the resulting quality measure axioms will also be unsatisfiable.

4 Axioms of clustering-quality measures

In this section we formalize the concept of a clustering-quality measure by proposing requirements ("axioms") that a clustering-quality measure should satisfy. We show that principles like those formulated in Kleinberg's axioms can be readily expressed in the framework of clustering-quality measures without leading to inconsistency. Although this gives a consistent set of requirements, some of them are too strong to require of all clustering-quality measures. We propose variants on these requirements, giving a consistent set of requirements that are satisfied by many natural quality measures.

4.1 Quality-measures versions of the Kleinberg axioms

A more lenient translation of Kleinber's axioms to the framework of quality-measures is by turning them into direct requirements about the relative values of the quality measure. Next, we illustrate that, under such a translation, Kleinberg's axioms can be readily expressed in our framework without resulting in inconsistency.

Definition 2 (Consistency) *Quality measure m satisfies* consistency *if for all clusterings C over (X, d), whenever d' is a C consistent variant of d, then $m(C, X, d) \geq m(C, X, d')$.*

Definition 3 (Scale Invariance) *Quality measure m satisfies* scale invariance *if for every clustering C of (X, d), and every positive λ, $m(C, X, d) = m(C, X, \lambda d)$.*

Definition 4 (Richness) *Quality measure m satisfies* richness *if for each non-trivial clustering C of X, there exists a distance function d over X such that $C = Argmax\{m(C, X, d)\}$.*

Consistency, scale invariance, and richness for clustering-quality measures form a consistent set of requirements. In particular, most quality measures presented in Section 5 (in particular, relative margin, additive margin, weakest link, and normalized cut) satisfy these three properties. In addition, consistency (although not richness) is satisfied by all the quality measures presented in Section 5.

4.2 Deriving axioms of quality measures

The previous section already offers a set of requirements for clustering-quality measures that capture the features that Kleinberg's axioms try to express by a satisfyable set of requirement. However, if one wishes to have a set of *axioms* for some domain, satisfiability is not enough. Axioms are meant to *define* the domain and, thus, should be such that every member of the domain satisfies all of them. We make the distinction between *properties* which are requirements satisfied by some clustering-quality measures and *axioms*. Only properties that should be satisfied by all objects in question should be called 'axioms'. We propose relaxations of the consistency and richness properties that can be used as axioms of quality measures. In addition, we present a new axiom, which captures an important property of clustering that is not modeled by Kleinberg axiom - the indifference of clusterings to the particular *identity* of the domain points. We end up with a consistent set of axioms for clustering-quality measures, satisfied by all the natural clustering quality measures we could come up with.

4.2.1 Isomorphism Invariance

We now introduce a new axiom, modeling the requirement that clustering should be indifferent to the individual identity of clustered elements. This axiom

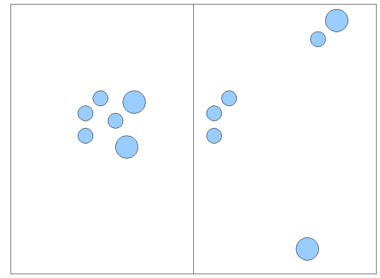

Figure 1: A consistent change of a 6-clustering.

of clustering-quality measures does not have a corresponding Kleinberg axiom. This axiom ensures that quality measures are independent of point description. That is, if the labels of all points are permuted, keeping the clustering fixed, the quality of the clustering should not change.

First, we need to define isomorphism between clusterings.

Definition 5 (Clustering Isomorphism) *Clusterings C and C' over (X, d) are* isomorphic*, $C \approx_d C'$, if there exists a distance-preserving isomorphism $\phi : X \to X$, such that $x \sim_C y$ if and only if $\phi(x) \sim_{C'} \phi(y)$.*

Definition 6 (Isomorphism Invariance) *Quality measure m is* isomorphism-invariant *for all clusterings C, C' over (X, d) where $C \approx_d C'$, $m(C, X, d) = m(C', X, d)$.*

We can readily reformulate this requirement in the framework of clustering functions. Namely, following a permutation on the data point's labels, the output of a clustering function should be isomorphic to the output prior to the permutation.

4.2.2 Consistency properties

We illustrate that consistency has some counterintuitive consequences. We then propose variants of consistency that overcome these shortcomings.

Consider the two data sets in Figure 1. The data set (X, d) on the left hand side of the figure has a good 6-clustering - call this clustering C. C is the most prominent clustering of (X, d). The data set on the right hand side is (X, d') where d' is a C-consistent variant of d. Notice that there is a 3-clustering on (X, d') which may have better quality than C on (X, d'). Therefore, a reasonable quality measure may assign better quality to C on (X, d) than to C on (X, d'). However, such quality measure would not satisfy the consistency property.

A simple modification of consistency, which we call local consistency, is the requirement that the distances between pairs of points within each cluster shrink uniformly, and distances between pairs of points in different clusters expand uniformly.

Definition 7 (Locally Consistent Variant) *Distance function d' is a C* locally consistent variant *of d, for a clustering C over (X, d), if*

- *For every cluster C_ℓ of C there is a constant $c_\ell \leq 1$, such that for all $x, y \in C_\ell$, $d'(x, y) = c_\ell d(x, y)$.*
- *There exists a $c \geq 1$ such that for every $x \not\sim_C y$, $d'(x, y) = c \cdot d(x, y)$.*

Definition 8 (Local Consistency) *Quality measure m is* locally consistent *if for all clusterings C over (X, d), whenever d' is a C locally consistent variant of d, then $m(C, X, d) \geq m(C, X, d')$.*

Local consistency has limited application in Euclidean spaces, where clustering often takes place. In Euclidean space, if we shrink each cluster uniformly, the distances between pairs of points in different clusters may change is a non-uniform manner.

Below we offer a more flexible version of consistency.

Definition 9 (Weakly Locally Consistent Variant) *Distance function d' is a C* weakly locally consistent variant *of d, where C is a clustering over (X, d), if*

- *For every cluster C_ℓ of C there is a constant $c_\ell \leq 1$, such that for all $x, y \in C_\ell$, $d'(x, y) = c_\ell d(x, y)$.*
- *For every $x \not\sim_C y$, $d'(x, y) \geq d(x, y)$.*
- *For some set of points containing a point p_ℓ from every cluster C_ℓ, there exists a constant $c \geq 1$ such that, for every $p_\ell, p_{\ell'}$, $d'(p_\ell, p_{\ell'}) = c \cdot d(p_\ell, p_{\ell'})$.*

Definition 10 (Weak Local Consistency) *Quality measure m is* weakly locally consistent *if for all clusterings C over (X, d), whenever d' is a C weakly locally consistent variant of d, then $m(C, X, d) \geq m(C, X, d')$.*

We propose weak local consistency as an axiom for clustering-quality measures. Note that weak local consistency implies consistency and local consistency.

4.3 Richness properties

Most of the quality measures presented in Section 5 satisfy richness, in particular, relative margin, additive margin, weakest link, and separability with the k-means loss function are rich.

However, richness is too strong a requirement for all quality measures. The reason for that is that many natural measures prefer more refined clusterings.

Definition 11 (Refinement) *A clustering C' of X is a* refinement *of clustering C of X if for every cluster C_i of C, there exists a set of clusters in C' that partition C_i.*

Definition 12 (Refinement Preference) *Quality measure m is* refinement-preferring *if for every clustering C of (X, d) that has a non-trivial refinement, there exists a non-trivial refinement C' of C such that $m(C', X, d) > m(C, X, d)$.*

Common loss functions such as k-means and k-median, as well as their L-Clustering Quality normalizations (presented in Section 5.1.1), satisfy the refinement preference property. For any refinement-preferring measure, given any clustering (that has a non-trivial refinement) over some data set, there is a non-trivial clustering of the data set with better quality. Thus, refinement-preferring measures are not rich.

We present a modification of quality measure richness, called co-final richness, which we propose as an axiom of quality measures. Co-final richness requires that by changing the distance function we can make the quality of any clustering arbitrarily good, however, we do not require that the quality of the clustering with the new distance function be better than the quality of all other clusterings with the new distance. On the other hand, we strengthen the richness requirement by specifying that the quality of the original clustering can be improved arbitrarily via consistent changes of the distance function.

Definition 13 (Co-final Richness) *Quality measure m satisfies* co-final richness *if for every pair of non-trivial clusterings C over (X, d) and C' over (X, d') there exists a C-consistent variant, d'', of d such that $m(C, X, d'') \geq m(C', X, d')$.*

4.3.1 Summary of axioms and properties

We proposed four axioms of clustering-quality measures: scale invariance, isomorphism invariance, weak local consistency, and co-final richness. It can be shown that all these axioms are necessary; that is, if we exclude any one of these axioms, there are functions that do not make good clustering-quality measures but satisfy the remaining axioms.

In addition, we discussed the following properties of quality measures: richness, consistency, and refinement preference.

5 Examples of clustering-quality measures

We propose clustering-quality measures for various common clustering paradigms; loss-based, center-based, and linkage-based clusterings. All our quality measures satisfy the axioms of clustering-quality measures presented in Section 4.

5.1 Loss-based quality measures

In this section, we present two quality measures for loss-based clustering. These quality measures, when used with common loss functions such as k-means or k-median, normalize these loss functions to obtain scale invariance while preserving other desirable properties (in particular the axioms presented in Section 4). We also discuss how spectral clustering relates to loss-based quality measures, and propose a quality measure based on spectral clustering theory.

A *clustering loss function* is a function $\mathcal{L} : \mathcal{C}_X \times D \to R^+ \cup \{0\}$, where $\mathcal{C}_X$ is the set of clusterings of data set X, and D is the family of distance functions over X. For example, the k-median loss function finds the sum of distances to the centers of clusters. For additional examples, consider functions of the form

$$\sum_{i=1}^{k} \frac{1}{|C_i|^\delta} \sum_{\{x,y\} \in C_i} d(x,y)^\beta$$

for $\delta, \beta \in \mathbf{R}$. For $\delta = 1$ and $\beta = 2$, we get the k-means loss function. For $\delta = 2$ and $\beta = 2$ the loss of a cluster is its variance. For $\delta = 0$ and $\beta = 1$ the loss of a cluster is the sum of all pairwise distances within the cluster.

5.1.1 $\mathcal{L}$-Clustering Quality

$\mathcal{L}$-Clustering Quality is a quality measure that normalizes clustering loss function $\mathcal{L}$. Let C_{all} denote the 1-clustering of X, that is, the clustering that groups all points in X into the same cluster.

Definition 14 ($\mathcal{L}$-Clustering Quality) *The* $\mathcal{L}$-Clustering Quality *of a clustering C over (X, d) is*

$$\mathcal{L}\text{-}CQ(C, X, d) = \frac{\mathcal{L}(C_{all}, X, d)}{\mathcal{L}(C, X, d)}.$$

Loss conformity: For loss-based quality measures, it can be desirable that a quality measure does not contradict a loss function. That is, when comparing two clusterings of a data set, we expect the clustering with lower loss to have better clustering quality. This requirement is not relevant for all clustering-quality measures, since often there is no relevant loss function. However loss conformity is desirable when the user believes that, modulo scale invariance, clusterings with lower loss are better. Whenever a quality measure satisfies this property for a clustering loss function $\mathcal{L}$, we say that it *conforms with* $\mathcal{L}$.

Definition 15 (Loss Conformity) *Quality measure m* conforms with $\mathcal{L}$ *if, for all C, C' and d over X, whenever $\mathcal{L}(C, X, d) \leq \mathcal{L}(C', X, d)$ then $m(C, X, d) \geq m(C', X, d)$.*

It is important to note that loss conformity addresses the behavior of a clustering-quality measure over a fixed distance function. Loss conformity is not relevant for comparing clusterings over different distance functions. We now show that $\mathcal{L}$-Clustering Quality conforms with $\mathcal{L}$.

Lemma 1 *$\mathcal{L}$-CQ conforms with $\mathcal{L}$.*

proof:

$$\mathcal{L}\text{-}CQ(C, X, d) = \frac{\mathcal{L}(C_{all}, X, d)}{\mathcal{L}(C, X, d)}.$$

Since $\mathcal{L}(C_{all}, X, d)$ is constant over all clusterings of (X, d), as $\mathcal{L}(C, X, d)$ decreases, $\mathcal{L}\text{-}CQ(C, X, d)$ increases. Therefore, given two clusterings of the same data set, the clustering with lower loss has better $\mathcal{L}\text{-}CQ$.

This property allows us to view $\mathcal{L}$-Clustering Quality as a normalized loss function; it preserves the relative order induced by the loss function over clusterings of the same distance function, while being scale-invariant.

Variance Ratio: A nice example of $\mathcal{L}$-Clustering Quality is for $\mathcal{L}(C, d) = avg_{x \sim_C y} d(x, y)$, the average distance within-cluster distance. This leads to a clustering-quality measure similar to a notion of clusterability by Zhang [6].

Let the *within-cluster variance* of a clustering C over (X, d) be $W(C, X, d) = avg_{x \sim_C y} d(x, y) = \frac{\sum_{x \sim_C y} d(x,y)}{|\{\{x,y\} \subseteq X | x \sim_C y\}|}$, the average distance between elements within the same cluster. Let the *between-cluster variance* of a clustering C over (X, d) be $B(C, X, d) = avg_{x \not\sim_C y} d(x, y) = \frac{\sum_{x \not\sim_C y} d(x,y)}{|\{\{x,y\} \subseteq X | x \not\sim_C y\}|}$, the average distance between elements in different clusters.

Definition 16 (Variance Ratio) *The* variance ratio *of a clustering C over (X, d) is*

$$VR(C, X, d) = \frac{B(C, X, d)}{W(C, X, d)}.$$

Note that $\mathcal{L}\text{-}CQ(C, d) = VR(C, d) + 1$, for $\mathcal{L}(C, d) = avg_{x \sim_C y} d(x, y)$. The range of variance ratio is $[0, \infty)$

and larger values of variance ratio indicate better clustering quality.

5.1.2 $\mathcal{L}$-Separability

We present an alternative method for loss function normalization, inspired by the notion of clusterability introduced by Ostrovsky et al. [5]

Let $C = \{C_1, C_2, \ldots, C_k\}$ be some k-clustering. For all $i \neq j$, let $C_{ij} = \{C \backslash \{C_i, C_j\} \cup \{C_i \cup C_j\}\}$ be a clustering identical to C, except with cluster C_i and C_j merged. We define separability as follows.

Definition 17 ($\mathcal{L}$-Separability) *The* $\mathcal{L}$-separability *of a clustering C over (X, d) is,*

$$\mathcal{L}\text{-}Sep(C, X, d) = \frac{\mathcal{L}(C, X, d)}{\min_{i,j} \mathcal{L}(C_{ij}, X, d)}.$$

Separability with loss functions such as k-means and k-median is sensitive to the minimal separation between clusters. If there are two clusters that are very close together, then the data set has low separability, regardless of how well separated are the rest of the clusters. Notice that variance ratio behaves differently on this aspect, since variance ratio can be made arbitrarily good by moving a single cluster far away from all the other clusters, regardless of the relationship between the other clusters.

Alternatively, we could define the quality measure

$$\mathcal{L}\text{-}Sep_{max}(C, X, d) = \frac{\mathcal{L}(C, X, d)}{\max\{\mathcal{L}(C_{ij}, X, d), i \neq j\}},$$

which is sensitive to the maximal separation between clusters. We could also choose to look at $\mathcal{L}\text{-}Sep_{avg}(C, X, s) = \frac{\mathcal{L}(C,X,d)}{avg\{\mathcal{L}(C_{ij},X,d)|i \neq j\}}$. Other variations we may consider involve merging more than two clusters of C.

5.1.3 Spectral clustering

We discuss how spectral clustering, a commonly-used clustering method, relates to loss-based clustering-quality measures. Spectral clustering is a relaxation of a graph cut problem. Consider the graph where the vertices V are the points in the data set and the edge weights represent a similarity function $s : V \times V \to \mathbf{R}^+$ between the points. Intuitively, clustering aims to separate the vertices into clusters such that the between-cluster edges have low similarity and the within-cluster edges have high similarity. In addition, spectral clustering requires that the cluster sizes be relatively balanced. This requirement is integrated in different ways for unnormalized and normalized spectral clustering.

Unnormalized spectral clustering aims to find clusterings that minimize a relaxation of the Ratio Cut (RC) function.

$$RC(C, X, s) = \sum_{x \not\sim y} s(x, y) \sum_{i=1}^{k} \frac{1}{|C_i|},$$

where $C = \{C_1, C_2, \ldots, C_k\}$ [3]. Normalized spectral clustering relaxes the Normalized Cut (NC) function [3].

$$NC(C, X, s) = \sum_{x \not\sim y} s(x, y) \sum_{i=1}^{k} \frac{1}{\sum_{\{x,y\} \subseteq C_i} s(x, y)}.$$

There are a number of ways to express similarity in terms of a distance function, one natural choice is to define similarity as $s(x, y) = \frac{1}{d(x,y)^2}$.

Ratio cut, like many other commonly-used loss functions, is not scale-invariant. However, it can be normalized using $\mathcal{L}$-separability (ratio cut cannot be used with $\mathcal{L}$-CQ, since the ratio cut of any 1-clustering is 0).

On the other hand, normalized cut can be used directly as a clustering-quality measure and satisfies all the axioms of clustering-quality measures presented in Section 4. The range of normalized cut is $[0, \infty)$ and lower values of normalized cut indicate better clustering quality.

5.2 Center-based quality measures

We now introduce quality measures for center-based clustering. We defined a center-based clustering as follows.

Definition 18 (Center-based Clustering) *A clustering $C = \{C_1, C_2, \ldots, C_k\}$ is* center-based *if there exist points, called* centers, *$c_1 \in C_1, c_2 \in C_2, \ldots, c_k \in C_k$, such that for all $x \in C_i$, $d(x, c_i) < d(x, c_j)$, for all $i \neq j$.*

That is, C is center-based if it is a Voronoi partition. Note that in our setting, where the input has no structure beyond a distance function, we cannot use the center of mass as the center of a cluster since it is not well-defined. Note that a center-based clustering is fully specified by its set of centers.

5.2.1 Relative Margin

For each point in the data set, consider the ratio of the distance from the point to its closest center, to the distance from the point to its second closest center.

Intuitively, when this ratio is smaller then a point is "more sure" to which cluster it belongs. We use the average ratio as a quality measure.

Definition 19 (Relative Point Margin) *The C-relative point margin of $x \in X$ is $C\text{-}RM_{X,d}(x) = \frac{d(x,c_i)}{d(x,c_j)}$, where c_i is the closest center to x, c_j is a second closest center to x, and C is a center-based clustering over (X, d).*

Definition 20 (Relative Margin) *The* relative margin *of a center-based clustering C over (X, d) is*

$$RM_{X,d}(C) = avg_{x \in X \setminus R} C\text{-}RM_{X,d}(x),$$

where R is the set of centers in C.

The range of relative margin is $[0, 1)$, and lower relative margin indicates a better clustering.

5.2.2 Additive Margin

We present an alternative approach for evaluating the quality of a center-based clustering. Instead of looking ratios, additive margin evaluates differences.

Definition 21 (Additive Point Margin) *The C-*additive point margin *of x is $C\text{-}AM_{X,d}(x) = d(x, c_j) - d(x, c_i)$, where c_i is the closest center to x, c_j is a second closest center to x, and C is a center-based clustering over (X, d).*

The additive margin of a clustering is the average additive point margin, divided by the average within-cluster distance. The normalization is necessary for scale invariance.

Definition 22 (Additive Margin) *The* additive margin *of a center-based clustering C over (X, d) is*

$$AM_{X,d}(C) = \frac{\frac{1}{|X|}\sum_{x \in X} C\text{-}AM_{X,d}(x)}{\frac{1}{|\{\{x,y\} \subseteq X | x \sim_C y\}|}\sum_{x \sim_C y} d(x, y)}.$$

The range of additive margin is $[0, \infty)$. Unlike relative margin, additive margin gives higher values to better clusterings.

5.3 Linkage-based quality measure

Linkage-based algorithms tend to look for a single tight path connecting points within each cluster, often finding very different clusterings than the ones found by common loss-based and center-based clustering algorithms. In linkage-based clustering, we can think of each pair of points from the same cluster as connected via a chain of points belonging to this cluster. The weakest link is the longest link over all such chains. The weakest link quality measure lets us evaluate the quality of clusterings detectable by linkage-based algorithms, such as single-linkage.

Definition 23 (Weakest Link Between Points) *The C-*Weakest Link *between points $x, y \in C_i$ is*

$$C\text{-}WL_{X,d}(x, y) = \min_{x_1, x_2, \ldots, x_\ell \in C_i} (\max(d(x, x_1), d(x_1, x_2), \ldots, d(x_\ell, y))),$$

where C is a clustering over (X, d) and C_i is a cluster in C.

The weakest link of C is the maximal value of $WL_{x \sim_C y}(x, y)$ over all pairs of points belonging to the same cluster, divided by the shortest between-cluster distance.

Definition 24 (Weakest Link of C) *The* Weakest Link *of a clustering C over (X, d) is*

$$WL(C) = \frac{\max_{x \sim y} C\text{-}WL_{X,d}(x, y)}{\min_{x \not\sim y} d(x, y)}.$$

The range of values of weakest link is $(0, \infty)$. Lower values of weakest link represent better clusterings.

5.4 Variants of quality measures

Given a clustering-quality measure, we can construct new quality measures with different characteristics by applying the quality measure on a subset of clusters. It suffices to consider a quality measure m that is defined for clusterings consisting of 2 clusters. Given such measure, we can create new quality measures. For example,

$$m_{min}(C, X, d) = \min_{S \subseteq C, |S|=2} m(S, X, d),$$

measures the worst quality of a pair of clusters in C.

Alternately, we can define, $m_{min}(C, X, d)$ and $m_{avg}(C, X, d)$, which evaluate the best or average quality of a pair of clusters in C. A nice feature of these variations is that if m satisfies the four axioms of clustering-quality measures then so do m_{min}, m_{max}, and m_{avg}.

6 Checking satisfiability of the axiom and properties

We analyze which axioms and properties are satisfied by the quality measures presented in Section 5. The requirement satisfaction displayed for $\mathcal{L}$-Clustering Quality and $\mathcal{L}$-separability apply for many different commonly-used loss functions; for example,

$\mathcal{L}$-separability and $\mathcal{L}$-Clustering Quality with the k-means loss function, and $\mathcal{L}$-Clustering Quality with the loss function used for variance ratio. Note that consistency implies local consistency.

	Loss			*Center*		*Linkage*
	$\mathcal{L}$-CQ	$\mathcal{L}$-Sep.	Nor. Cut	Rel Margin	Add Margin	Weakest Link
Properties						
Richness	-	+	+	+	+	+
K-Consist.	+	-	+	+	-	-
Consistency	+	+	+	+	+	+
Refinement Preference	+	-	-	-	-	-
Axioms						
Scale Invariance	+	+	+	+	+	+
Iso. Invariance	+	+	+	+	+	+
Weak Local Cons.	+	+	+	+	+	+
Co-final Richness	+	+	+	+	+	+

Proof: Omitted due to lack of space.

7 Computational complexity

For a clustering-quality measure to be useful, it is important to be able to quickly compute the quality of a clustering using that measure. The quality of a clustering using the measures presented in Section 5 can be computed in low polynomial time, in terms of n (the number of points in the data set) and k (the number of clusters in the clustering).

We can find the relative point margin of all points in $O(nk)$, therefore the relative margin of a given clustering can be computed in $O(nk)$. It is also easy to show that additive margin can be found in $O(n^2)$ operations and the weakest link of a clustering can be computed in $O(n^3)$ operations.

Assuming that the loss function $\mathcal{L}$ of a k-clustering can be evaluated in $g(n,k)$ operations where g is a polynomial function, we can find $\mathcal{L}$-separability and $\mathcal{L}$-Clustering Quality in polynomial time. Since $|\{C_{ij} \mid i \neq j\}| = \binom{k}{2}$, we can find $\mathcal{L}$-$Sep(C,d)$ in $\binom{k}{2}g(n,k-1)+g(n,k)$ operations. Similarly, $\mathcal{L}$-$Sep(C,d)$ can be found in $g(n,1)+g(n,k)$ operations. For instance, if $\mathcal{L}$ is the k-means loss function, which can be evaluated in $O(n^2)$ operations, then we can find k-$means$-$Sep(C,d)$ in $O(k^2n^2)$ operations and k-$means$-CQ in $O(n^2)$ operations. Note also that the normalized cut of a given clustering can be computed in $O(n^2)$ operations.

References

[1] N. Jardine, and R. Sibson. *Mathematical Taxonomy Wiley*, 1971.

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

[3] U. von. Luxburg. (2006, August). “A tutorial on spectral clustering.” In (MPITechnical Reports No. 149). Tubingen: Max Planck Institute for Biological Cybernetics.

[4] J. Puzicha, T. Hofmann, and J. Buhmann. “Theory of Proximity Based Clustering: Structure Detection by Optimization,” Pattern Recognition, 33(2000).

[5] R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. “The Effectiveness of Lloyd-Type Methods for the k-Means Problem.” *Foundations of Computer Science*, 2006. FOCS ’05. 47th Annual IEEE Symposium. Berkeley, CA, USA, Oct. 2006. pp. 165-176.

[6] Bin Zhang. “Dependence of Clustering Algorithm Performance on Clustered-ness of Data.” Technical Report, 20010417. Hewlett-Packard Labs, 2001.