

A Theoretical Study of Clusterability and Clustering Quality

by

Margareta Ackerman

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2007

©Margareta Ackerman 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Clustering is a widely used technique, with applications ranging from data mining, bioinformatics and image analysis to marketing, psychology, and city planning. Despite the practical importance of clustering, there is very limited theoretical analysis of the topic. We make a step towards building theoretical foundations for clustering by carrying out an abstract analysis of two central concepts in clustering; clusterability and clustering quality.

We compare a number of notions of clusterability found in the literature. While all these notions attempt to measure the same property, and all appear to be reasonable, we show that they are pairwise inconsistent. In addition, we give the first computational complexity analysis of a few notions of clusterability.

In the second part of the thesis, we discuss how the quality of a given clustering can be defined (and measured). Users often need to compare the quality of clusterings obtained by different methods. Perhaps more importantly, users need to determine whether a given clustering is sufficiently good for being used in further data mining analysis. We analyze what a measure of clustering quality should look like. We do that by introducing a set of requirements ('axioms') of clustering quality measures. We propose a number of clustering quality measures that satisfy these requirements.

Acknowledgements

I would like to thank my supervisor, Shai Ben-David, for his help with this thesis and the numerous fruitful discussions. I would also like to thank him for introducing me to clustering, and inspiring an interest and fascination with the subject. I would like to thank David Loker, for proofreading this thesis at various stages, and for his sharp observations and excellent suggestions. I would like to thank R. Wayne Oldford and Ming Li for their useful comments and suggestions. I would also like to thank Daniil Golod, for proofreading this thesis and for his interesting suggestions. A thanks goes to David Pal, for one particularly useful discussion on clusterability. Lastly, I would like to thank my parents, Efim Ackerman and Anna Dolinsky - for believing in me.

Contents

1	Introduction	1
1.1	Preliminaries	2
2	Clusterability	4
2.1	Introduction	4
2.2	Notions of Clusterability from the Literature	5
2.2.1	Separability Clusterability	6
2.2.2	Worst Pair Ratio Clusterability	7
2.2.3	Variance Ratio Clusterability	10
2.3	Perturbation-Based Notions of Clusterability	13
2.4	Computational Complexity of Clusterability	17
2.4.1	Computational Complexity of Separability	17
2.4.2	Computational Complexity of Variance Ratio	18
2.4.3	Computational Complexity of Worst Pair Ratio	19
2.5	Comparisons of Notions of Clusterability	22
2.5.1	Separability versus Variance Ratio	23
2.5.2	Worst Pair Ratio versus Variance Ratio	27
2.5.3	Worst Pair Ratio versus Separability	30
2.5.4	Perturbation Loss versus Worst Pair Ratio	30
2.5.5	Perturbation Loss versus Variance Ratio	32
2.5.6	Perturbation Loss versus Separability	36
2.6	Conclusions	38
3	Clustering Quality	40
3.1	Introduction	40
3.2	Previous Work	41
3.2.1	Clustering Functions	41
3.2.2	Clusterability	43
3.3	Axioms of Clustering Quality Measures	44
3.3.1	Axioms Non-redundancy	47
3.4	Examples of Quality Measures Satisfying the Axioms	49
3.4.1	Variance Ratio	50
3.4.2	Separability	54

3.4.3	Margins	60
3.5	Minimal Subset Quality	64
3.6	Clusterability and Clustering Quality	65
3.7	Conclusions	66
4	Conclusions	67

Chapter 1

Introduction

Clustering is a widely used technique, with applications ranging from data mining, bioinformatics and image analysis to marketing, psychology, and city planning. Clustering is the problem of identifying groups of similar objects. Despite the practical importance of clustering, there is very limited theoretical analysis of the topic. We make a step towards building theoretical foundations for clustering by carrying out an abstract analysis of two central concepts in clustering; clusterability and clustering quality.

Clustering methods can be categorized into algorithm-based and loss-based. Algorithm-based clustering methods find clusterings using a specific algorithm. Linkage-based and spectral clustering methods fall under this category. Loss-based clustering methods define the optimal clusterings of a data set as the clusterings that minimize some loss function.

We compare a number of notions of clusterability found in the literature. Most of these notions of clusterability apply to loss-based clustering. While all these notions attempt to measure the same property, and all appear to be reasonable, we show that they are pairwise inconsistent. Understanding how these notions relate to one another enables the formal comparison of results on clusterability. In addition, understanding the relationships between different notions of clusterability, such as their relative strength, helps to make informed

choices when selecting clusterability notions for particular studies. We also give the first computational complexity analysis of a few notions of clusterability.

We introduce a new notion of clusterability based on point perturbations. This notion of clusterability is well-suited for center-based clustering over normed vector spaces. We compare this notion to the notions of clusterability from the literature, finding that it is distinct from these notions. We also present variations of this notion, which opens up a whole new class of notions of clusterability.

In the second part of the thesis, we discuss how the quality of a given clustering can be defined (and measured). Users often need to compare the quality of clusterings obtained by different methods. Perhaps more importantly, users need to determine whether a given clustering is sufficiently good for being used in further data mining analysis. We analyze what a measure of clustering quality should look like. We introduce a set of axioms of clustering quality measures. We then propose a number of clustering quality measures that satisfy these axioms. These clustering quality measures apply in different settings and have different properties. We introduce a number of measures specifically for loss-based clustering, as well as a number of measures for center-based clustering.

In this thesis we take the first step towards establishing a theory of clustering. Throughout this thesis, we propose alternative formalizations of clustering quality measures and some notions of clusterability. We also discuss connections between clusterability and clustering quality. We hope that this work will lead to further development of the theory of clustering and believe that it will be of great benefit in practical applications.

1.1 Preliminaries

A *k*-clustering of data set X is a k -partition of X , that is, a set of k disjoint subsets of X such that their union is X . A *clustering* of X is a k -clustering of X for some $k \geq 1$. A

clustering is *trivial* if each cluster consists of a single point, or if all points belong to the same cluster. A clustering is *non-trivial* if it is not trivial. For $x, y \in X$ and clustering C of X , $x \sim_C y$ whenever x and y are in the same cluster of clustering C and $x \not\sim_C y$, otherwise. We introduce additional definitions throughout the thesis, as appropriate.

Chapter 2

Clusterability

2.1 Introduction

Clusterability is a central concept in clustering. The goal of clustering is to find meaningful patterns in data and a measure of clusterability determines to what degree these patterns exist. Authors often come up with new definitions of clusterability depending on the use of clustering in their research. We selected a number of clusterability notions from the literature. While all these notions attempt to measure the same property, and all appear to be reasonable definitions, we found that they are pairwise inconsistent. Understanding how these notions relate to one another enables the formal comparison of results on clusterability. In addition, understanding the relationships between different notions of clusterability, such as their relative strength, helps to make informed choices when selecting clusterability notions for particular studies.

The other dimension on which we analyze notions of clusterability is so natural and important that it is surprising that it has not yet been studied. We ask the following question: For a given notion of clusterability, how hard is it to determine the level of clusterability of a data set? The answer to this question is central to the practical value of the notion in ques-

tion. We give the first computational complexity analysis of a few notions of clusterability.

We also introduce a new notion of clusterability based on point perturbations. This notion of clusterability is well-suited for center-based clustering and works on data sets over normed vector spaces. We compare this notion to the notions of clusterability from the literature, finding that it is distinct from these notions. We also present variations of this notion, which opens up a whole new class of notions of clusterability.

In this chapter we work with data sets over normed vector spaces. Since the notions of clusterability taken from the literature were originally defined for Euclidean spaces, our definitions are more general. However, all definitions and most results pertaining to the notions of clusterability from the literature can be easily generalized further, to arbitrary spaces (where the input is a set of pairwise distances between points).

We now give an outline of this chapter. First, we introduce notions of clusterability from the literature. We then introduce our new notion of clusterability as well as its variations. Next we prove computational complexity results for the notions of clusterability appearing in the literature. Then we perform a pairwise comparison of the notions and conclude with open problems.

2.2 Notions of Clusterability from the Literature

A *notion of clusterability* is a function that takes a data set X and returns a real value. In this chapter, we assume that the data sets are over normed vector spaces, therefore distances between points are implicitly specified by the data sets. Since many of the notions found in the literature are defined with respect to the k -means loss function, we focus our analysis on clustering with respect to k -means. However, the notions of clusterability presented in this chapter generalize in a natural way to work with other loss functions.

Given a data set X , the k -means problem is to find a k -partition $\{X_1, X_2, \dots, X_k\}$ of

X such that the k -means loss function $\sum_{i=1}^k \sum_{x \in X_i} \|x - c_i\|^2$ is minimized, where $c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$ is the center of mass of X_i . In this chapter, an optimal clustering refers to a clustering with minimal k -means loss, unless specified otherwise. We let $OPT_k(X)$ denote the loss of an optimal k -means clustering of X .

We introduce three notions of clusterability that appear in the literature. For each notion, we discuss its range of values, previous work, and some relevant properties.

2.2.1 Separability Clusterability

In their paper “The Effectiveness of Lloyd-Type Methods for the k -Means Problem,” Ostrovsky, Rabani, Schulman, and Swamy define the notion of ϵ -separability[12].

Definition 1 (Separability). *A data set X is (k, ϵ) -separable if $OPT_k(X) \leq \epsilon OPT_{k-1}(X)$.*

For convenience, we define $S_k(X)$ to be the smallest ϵ such that X is (k, ϵ) -separable. A data set has better separability than another data set if it is separable for smaller ϵ . Data is considered well-clusterable by the separability notion of clusterability if it is ϵ -separable for sufficiently small ϵ .

Range of values

Let’s consider the upper and lower bounds of separability. Separability can be as good as $\epsilon = 0$. In particular, for any k , there is a data set that is $(k, 0)$ -separable. An example of such a data set is a set of k points positioned in different locations. Then $OPT_k(X) = 0$ and $OPT_{k-1}(X) > 0$.

How high can ϵ be? Since the loss of an optimal k -clustering is no worse than the loss of an optimal $(k - 1)$ -clustering, $S_k(X) \leq 1$. We show that separability can be arbitrarily close to 1. Consider a data set X uniformly distributed on a line segment

of length L . Then $OPT_k(X) \approx n \frac{L^2}{16k^2}$ for large enough n . Therefore, $S_k(X) \approx (\frac{k-1}{k})^2$. Therefore, as k goes to infinity, ϵ approaches 1.

Relevant Results in Literature

The Lloyd algorithm is a simple center-based clustering algorithm. Initially, select k centers. Then, perform a Lloyd step: find a k -partition by assigning each point to its closest center, and let the new centers be the centers of mass of the resulting clusters. Perform Lloyd steps until there is a step that does not change the centers. Many variations of the Lloyd algorithm appear in the literature. The original greedy iteration to minimize loss was proposed by Lloyd [8].

Ostrovsky et al. show that for data sets with better separability, a modified version of the Lloyd algorithm yields solutions of better quality with higher probability[12]. An interesting feature of this notion of clusterability is that it implies and is implied by the condition that two near optimal solutions differ by a small fraction of data points, as shown by Ostrovsky et al.[12]

2.2.2 Worst Pair Ratio Clusterability

Worst Pair Ratio (WPR) is a measure of clusterability that is based on distances between points that are within and between clusters. Given clustering C , we denote the minimum distance between two points in different clusters of C as the *split* between the two clusters, and the minimal split between two clusters as the split of C ; that is, $split_C(X) = \min_{x \not\sim y} \|x - y\|$. We denote the maximum distance between two points within a cluster in C as the *width* of the cluster, and the maximal width of a cluster in C as the width of C ; $width_C(X) = \max_{x \sim y} \|x - y\|$.

We focus on optimal clusterings by the k -means measure of optimality. However, we

could use any other measure of optimality.

Definition 2 (Worst Pair Ratio). *Let \mathcal{C} be the set of optimal k -means clusterings of X . Then the worst pair ratio of X for k is*

$$WPR_k(X) = \max_{\mathcal{C} \in \mathcal{C}} \frac{\text{split}_{\mathcal{C}}(X)}{\text{width}_{\mathcal{C}}(X)}.$$

Note that in most real data sets, there is a unique clustering with minimal k -means loss. An alternative natural definition of WPR is the maximum ratio of split over width over all k -clusterings of the data set. This version has been presented by Epter et al.[5] The definition that we have chosen has the same flavour as the other notions presented in this chapter. However, since these definitions are very similar, the complexity and comparison results presented here apply to Epter’s version of WPR as well as the version that we focus on, using essentially the same proofs.

The higher the value of $WPR_k(X)$, the more clusterable is the data set. We call data set X WPR well-clusterable for k when $WPR_k(X) > 1$ and WPR poorly-clusterable for k , otherwise.

Range of values

How high can WPR be? For $k = 1$, between-cluster variance is poorly defined; we thus consider the problem for $k \geq 2$. We can easily see that WPR can be arbitrarily large. For instance, consider k well spaced out clusters in \mathbf{R} . Then by uniformly spacing out the clusters, we can increase WPR arbitrarily.

How low can WPR be? Consider a set of n uniformly distributed points in \mathbf{R} . The optimal 2-means clustering groups the leftmost half of the points into one cluster and all the other points into another cluster. By increasing the number of points in this scenario, we can get values of WPR arbitrarily close to 0. To extend this example to

arbitrary values of k , add $k - 2$ points sufficiently far from all other points and from each other so that they make their own clusters in the optimal k -means clustering. Thus, the range of WPR for $k \geq 2$ is $(0, \infty)$.

Multiple Optimal Solutions

We illustrate that different optimal clusterings of the same data set may have different minimum split, maximum cluster width, and split over width values. Consider the points $1, 3, 14$ and $14 + 8\sqrt{3}$ in one-dimensional space. We cluster this data into 2 clusters. One optimal 2-means clustering is $\{A, B\}$ where $A = \{1, 3, 14\}$ and $B = \{14 + 8\sqrt{3}\}$. The center of cluster A is 6 and the center of cluster B is $14 + 8\sqrt{3}$. Therefore, the loss of this clustering is $5^2 + 3^2 + 8^2 + 0 = 98$. Another clustering has clusters $C = \{1, 3\}$ and $D = \{14, 14 + 8\sqrt{3}\}$. The center of C is 2 and the center of D is $14 + 4\sqrt{3}$. Therefore, the loss of the second clustering is $1 + 1 + (4\sqrt{3})^2 + (4\sqrt{3})^2 = 98$. Clearly, there is no 2-clustering of this set with loss less than 98. Therefore, we have two optimal clusterings, one with minimum separation $8\sqrt{3} \approx 13.9$ and maximum cluster width 13 and the other minimum separation 11 and maximum cluster width $8\sqrt{3} \approx 13.9$. So we have split over width of ≈ 1.065 in one of the optimal clusterings, and ≈ 0.794 in the other. So, according to one of the optimal clusterings the data is well-clusterable and according to the other it is not. Therefore, we select the optimal clustering with maximal split over width ratio in our definition of WPR.

Sensitivity to Noise and Outliers

One of the shortcomings of this notion is its intolerance to noise and outliers. For instance, having only one out of a thousand clusters with high width pulls down the worst pair ratio, suggesting that what may intuitively be well-clusterable data, is poorly-clusterable by this notion of clusterability. Similarly, having a small num-

ber of non-representative points can decrease the separation, labeling data that has “innate” clusters as poorly-clusterable.

When using WPR as a measure of clusterability on real data sets it may be helpful to preprocess data by removing noise and outliers. One way to remove noise is to run an algorithm which detects dense areas, and then removes all points in areas of density below a certain threshold. On the other hand, we could use WPR to detect noise, by allowing the removal of a preset number of points that pulls down the worst pair ratio.

Without preprocessing, this is a rather unforgiving notion, but it is simple and has nice properties. It is satisfying in a least one direction - if according to this notion a data set is well clusterable, then it does have what seems to be an innate clustering. On the other hand, if according to WPR a data set is poorly-clusterable, it is still possible that it possesses an innate clustering that is not captured by this notion.

Relevant Previous Work

Epter et al.[5] present a heuristic for finding the number of clusters to use to cluster data set X , assuming there is some clustering of the data set where the maximal width of a cluster is smaller than the minimal split of the clustering. They present a method for estimating a suitable number of clusters for X and a threshold c , so that two points are in the same cluster if and only if they are at distance less than c .

2.2.3 Variance Ratio Clusterability

Variance Ratio (VR) is a very natural notion of clusterability that measures the ratio of the between-cluster variance over the within-cluster variance. This measure can be viewed as a relaxation of WPR. We can define the variance ratio of a data set X as $\frac{avg_{x \not\sim_C y} \|x-y\|}{avg_{x \sim_C y} \|x-y\|}$, where C is an optimal clustering by some measure of optimality.

A specialized version of this idea was presented by Bin Zhang [13]. Zhang presents a version of variance ratio that is tailored for the k -means loss function. Recall that the variance of X is $\sigma^2(X) = \frac{1}{|X|} \sum_{x \in X} \|x - c\|^2$, where c is the center of mass of X . Let $C = \{X_1, X_2, \dots, X_k\}$ be a clustering, where $center(X_i) = c_i$. Let $p_i = \frac{|X_i|}{|X|}$. Let $B_C(X) = \sum_{i=1}^k p_i \|c_i - c\|^2$ denote the between-cluster variance of C . Let $W_C(X) = \sum_{i=1}^k p_i \sigma^2(X_i)$ denote the within-cluster variance of C . Then,

$$\sigma^2(X) = W_C(X) + B_C(X) = \sum_{i=1}^k p_i \sigma^2(X_i) + \sum_{i=1}^k p_i \|c_i - c\|^2.$$

Notice that $W_C(X)$ is the loss function that k -means minimizes divided by $|X|$.

Definition 3 (Variance Ratio). *Let \mathcal{C} be the set of optimal k -mean clusterings of data set X . The variance ratio of X for k is*

$$VR_k(X) = \max_{C \in \mathcal{C}} \frac{B_C(X)}{W_C(X)}.$$

Since $\sigma^2(X) = W_C(X) + B_C(X)$ and $\sigma^2(X)$ is constant over all clustering of X , $W_C(X)$ as well as $B_C(X)$ are equal in all optimal k -means clusterings that maximize the between over within variance ratio. Thus, we let $W_k(X) = W_C(X)$ and $B_k(X) = B_C(X)$, where C is some optimal k -means clustering that maximizes the between over within variance ratio over all optimal k -means clusterings of X . Note that for higher values of VR, data is better clusterable.

Range of Values

What is the maximum value of VR clusterability? For $k = 1$, $VR_1(X) = 0$, since $c_1 = c$. For $k \geq 2$, $VR_k(X)$ can be arbitrarily large, by moving a single point arbitrarily far away from the others. Note the sensitivity of the VR measure to the positioning of a

single outlier. By placing a single point far away, it is possible to arbitrarily increase $VR_k(X)$.

How low can $VR_k(X)$ be for $k \geq 2$? Clearly, since distances are non-negative, $VR_k(X) \geq 0$. It remains open to find the lower bound of VR.

Relevant Results in Literature

Zhang [13] found through experiments that the more clusterable data is according to VR clusterability, the harder it is to cluster. In particular, Zhang analyzes the quality of algorithms for the following three problems: k -means, expectation minimization with linear mixing of the Gaussian density function, and k -harmonic means. Let M_{loc} be a local optimal clustering solution obtained by some algorithm. The Quality Ratio is defined as

$$QR = \sqrt{k\text{-means}(X, M_{loc})} / \min_M \sqrt{k\text{-means}(X, M)}.$$

That is, the quality ratio is the k -means loss using centers M_{loc} over the optimal k -means loss. M_{loc} could have been obtained using algorithms for any of the three listed problems. If QR is low, the quality of a clustering is better. The average QR is analyzed as a function of data set clusterability.

Zhang finds that for all three algorithms, the quality ratio is higher when the data is better clusterable. That is, as $VR_k(X)$ becomes larger, the loss obtained by the algorithms differs more from the loss of the optimal solution. The explanation proposed by Zhang for this phenomenon is that for well-clusterable data, incorrectly placed centers incur higher penalty on the loss function.

2.3 Perturbation-Based Notions of Clusterability

We now introduce a new notion of clusterability. There are a number of ways to define clusterability in terms of perturbation. We present a version that is particularly well-suited for center-based clustering. In the end of the section, we present a number of alternative formulations. Given a good center-based clustering, if the centers of an optimal solution are perturbed slightly, yielding a new clustering, the loss of the resulting clustering should be close to the loss of the original clustering. Since we are working with normed vector spaces, we define the center of a cluster as its center of mass. That is, cluster X_i has center

$$c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x.$$

We give a definition of perturbation loss with the k -means loss function, although any loss function can be used. First, we need a preliminary definition.

Definition 4 (ϵ -close). *Let X be a data set over S . Let $C = \{X_1, X_2, \dots, X_k\}$ be a clustering, with centers $\{c_1, c_2, \dots, c_k\}$. Clustering $C' = \{X'_1, X'_2, \dots, X'_k\}$ of X is ϵ -close to C if there exists a set $\{c'_1, c'_2, \dots, c'_k\} \subseteq S$ such that $\|c'_i - c_i\| \leq \epsilon$, and whenever $x \in X'_i$ then $\|x - c'_i\| \leq \|x - c'_j\|$ for all $i \neq j$.*

Note that the loss of C' is computed with respect to the centers of mass of its clusters, which may not necessarily be $\{c'_1, c'_2, \dots, c'_k\}$.

Definition 5 (Perturbation Loss Clusterability). *Let ϵ be a non-negative real number, and $f : \mathbf{R}^+ \cup \{0\} \rightarrow \{r \in \mathbf{R}, r \geq 1\}$ a monotonic non-decreasing function. X is (ϵ, f) -PL clusterable for k if for any k -means optimal clustering C , and C' ϵ -close to C , $k\text{-means}(C') \leq f(\epsilon)k\text{-means}(C)$.*

Thus, if f is a slow-growing function and ϵ is reasonably large relative to the positions of points in the data set, then X is well-clusterable.

An application of PL clusterability

We present an algorithm for finding a clustering that is ϵ -close to a k -means optimal clustering. If we know that X is (ϵ, f) -PL clusterable, then we can estimate the quality of the clustering found by the algorithm.

The following algorithm is based on an algorithm by Ben-David, Eiron, and Simon [3]. The algorithm checks all possible sets of k sets each having l points from X , where the average of each set of l points defines a center. For each set of k centers, it performs a Lloyd step to obtain a clustering. In particular, when the initial set of centers is found, it finds the closest center for each point and recalculates the centers of the resulting clusters. It chooses the clustering that gives the best k -means loss of the clusterings found using this approach.

Algorithm 1.

INPUT: A data set X , $k \geq 1$, $l \geq 1$.

OUTPUT: Outputs a clustering C_A of X such that

$k\text{-means}(C_A) \leq \min \{k\text{-means}(C) \mid C \text{ is } \frac{R}{\sqrt{l}}\text{-close to an optimal clustering.}\}$ where

R is the radius of the minimum hypersphere that contains all the points in X .

$C_A = \emptyset$

for each set of k l -subsets of X

 find the center of mass for each l -subset

 for each point in X , find the closest of these centers, getting clustering \hat{C}

 let \hat{C}' be the set of the centers of mass of \hat{C}

 calculate the loss of clustering with centers in \hat{C}'

 if $(C_A = \emptyset \text{ or } k\text{-means}(C_A) > k\text{-means}(\hat{C}'))$

$C_A = \hat{C}'$

return C_A

We now show how the above algorithm relates to PL clusterability.

Lemma 1. *Given integers $l, k \geq 1$ and data set $X \subseteq \mathbf{R}^m$ that is $(\frac{R}{\sqrt{l}}, f)$ -PL clusterable, Algorithm 1 finds k -clustering C_A such that $k\text{-means}(C_A) \leq f(\frac{R}{\sqrt{l}})k\text{-means}(C)$, where C is an optimal k -means clustering of X .*

Proof. Note that the radius of the data set, R , can be calculated by finding the points of maximum distance and dividing by 2. The following is a result by Maurey [10]: For any fixed $l \geq 1$ and each x' in the convex hull of X , there exist $x_1, x_2, \dots, x_l \in X$ such that $\|x' - \frac{1}{l} \sum_{i=1}^l x_i\| \leq \frac{R}{\sqrt{l}}$ where R is the radius of the smallest hypersphere that contains all of the points in X . Therefore, there is a clustering, \hat{C} , examined by Algorithm 1, that is $\frac{R}{\sqrt{l}}$ -close to the optimal clustering. After the Lloyd step, yielding clustering \hat{C}' , the loss does not decrease. Since Algorithm 1 selects the minimal loss clustering of the ones it reviews, $k\text{-means}(C_A) \leq k\text{-means}(\hat{C}')$. Since \hat{C} is $\frac{R}{\sqrt{l}}$ -close to C , $k\text{-means}(C_A) \leq k\text{-means}(\hat{C}) \leq f(\frac{R}{\sqrt{l}})k\text{-means}(C)$. \square

By Lemma 1, if we know that X is $(\frac{R}{\sqrt{l}}, f)$ -PL clusterable, then we have a quality guarantee on the output of Algorithm 1. Note that for sufficiently large values of l , $k\text{-means}(C_A) = k\text{-means}(C)$.

The running time of Algorithm 1 is mn^{lk+1} , where $n = |X|$ and $X = \mathbf{R}^m$. We can remove the dependence of the running time on $|X|$ by sampling. Ben-David [2] showed that if a sample $S \subseteq X$ of size $\geq \frac{\ln 2/\delta}{2\gamma^2}$ is picked i.i.d then with probability $> 1 - \delta$ the centers found using S as the data set extend to a clustering of X of loss less than γ away from the optimal clustering of X . Therefore, by using a sample of the data set, we remove the dependence of Algorithm 1 from the size of the data set. Clearly, this assumes that the data set is sufficiently large; however, this is permissible since for small data sets the k -means problem is easier to solve exactly. Depending on the level of certainty required, select a big enough sample. The running time of A with the use of a sample is then $m|S|^{lk+1}$.

Alternative Formulations

Instead of comparing loss function values, we could consider the distance between the optimal clustering and the clustering resulting from slight perturbation of the centers. There are many ways to define the distance between clusterings, see [9] for a review of some examples.

For instance, the following is a definition of distance between two clusterings \hat{C} and C' of data set X as defined by Meila [9]. Let $\hat{C} = \{\hat{C}_1, \hat{C}_2, \dots, \hat{C}_k\}$ and $C' = \{C'_1, C'_2, \dots, C'_k\}$. Let $m_{i,j} = |\hat{C}_i \cap C'_j|$. Then the distance between clusterings \hat{C} and C' is $\mathcal{D}(\hat{C}, C') = 1 - \frac{1}{n} \max_{\{\pi \in \Pi\}} \sum_i m_{i, \pi(i)}$, where n is the number of points in the data set and Π is the set of all permutations of $\{1, 2, \dots, k\}$. We define Perturbation Distance clusterability using Meila's definition of cluster distance, however, the definition can work with many other ways of defining distances between clusterings.

Definition 6 (Perturbation Distance Clusterability). *Let ϵ be a non-negative real number, and $f : \mathbf{R}^+ \cup \{0\} \rightarrow \mathbf{R}^+ \cup \{0\}$ a function. X is (ϵ, f) -PD clusterable for k if for any k -means optimal clustering C , and C' an ϵ -close clustering to C , $\mathcal{D}(C, C') \leq f(\epsilon)$.*

We now present a different variation of perturbation-based clusterability. In soft clustering, a single point is assigned to many clusters with various probabilities. In hard clustering, which is the type of clustering with which we are concerned, a point is assigned to a unique center (or the cluster which this center represents). Sometimes there is a unique center that is by far the most appropriate for a specific point, whereas in other situations many centers are about equally well-suited for a point. We can think of the difference between the distance to the closest center and the distance to the second closest center as the margin of a point. Then, we use the average margin as a notion of clusterability. We now formalize these ideas.

Definition 7 (Relative Point Margin). *Let C with cluster centers $\{c_1, c_2, \dots, c_k\}$ be a clustering. Let $x \in X$. Let c_i be the closest center to x and c_j the second closest center to x . The relative margin of x with respect to C is $RM_C(x) = \frac{\|x - c_i\|}{\|x - c_j\|}$.*

Definition 8 (Relative Margin). *Let \mathcal{C} be the set of optimal k -means clusterings of X . The relative margin of X is*

$$RM_k(X) = \min_{C \in \mathcal{C}} \text{avg}_{x \in X} RM_C(x).$$

2.4 Computational Complexity of Clusterability

The computational complexity of clusterability has received surprising little attention in the literature. We present the first results on the computational complexity of the notions of clusterability appearing in the literature that were presented in Section 2.2.

2.4.1 Computational Complexity of Separability

Separability and Variance Ratio were originally presented for data sets over Euclidean spaces. We show that over Euclidean spaces, determining the separability and VR of a data set is NP-hard.

Theorem 1. *Given $X \subseteq R^m$, integer $k \geq 2$, and $0 \leq \epsilon < 1$, it is NP-hard to determine whether $S_k(X) \leq \epsilon$.*

Proof. The decision version of the k -means problem is: does there exist a set of centers such that the loss function of k -means has value $\leq v$ if these centers are used to cluster X ? This problem is NP-complete for $k \geq 2$ over Euclidean spaces.[4]

X is $(2, \epsilon)$ -separable clusterable if $\frac{OPT_2(X)}{OPT_1(X)} \leq \epsilon$. Suppose that we can determine whether $X \subseteq R^m$ is $(2, \epsilon)$ -separable for any arbitrary constant $\epsilon > 0$ in polynomial-time. Then since

$OPT_2(X) \leq \epsilon OPT_1(X)$ and $OPT_1(X)$ can be found in polynomial-time, we can find out if $OPT_2(X) \leq \mu$ for any μ by checking if X is $(2, \epsilon)$ -separable for $\epsilon = \frac{\mu}{OPT_1(X)}$. However, determining if $OPT_2(X) \leq \mu$ for any arbitrary $\mu > 0$ is NP-hard. Therefore, determining if X is $(2, \epsilon)$ -separable is NP-hard.

Since the problem is NP-hard for $k = 2$, it is NP-hard for $k > 2$. To show that the problem is NP-hard for $k > 2$, we reduce the problem for $k = 2$ to the problem for any $k > 2$. Given X , add $k - 2$ points sufficiently far away from all points in X and from each other, so that each one of the new points will have to be its own cluster in the optimal clustering. Then in any k -means optimal clustering, the remaining 2 clusters are an optimal 2-means solution for the original data set. \square

2.4.2 Computational Complexity of Variance Ratio

Theorem 2. *Given $X \subseteq R^m$ and integers $k \geq 2$ and $r \geq 0$, it is NP-hard to determine whether $VR_k(X) \geq r$.*

Proof. We know that $\sigma^2(X) = W_k(X) + B_k(X)$. So,

$$\begin{aligned} VR_k(X) &= \frac{B_k(X)}{W_k(X)} \\ &= \frac{\sigma^2(X) - W_k(X)}{W_k(X)} \\ &= \frac{\sigma^2(X)}{W_k(X)} - 1 \\ &= \frac{|X|\sigma^2(X)}{OPT_k(X)} - 1 \end{aligned}$$

Thus, if we can tell whether $VR_k(X) = \frac{|X|\sigma^2(X)}{OPT_k(X)} - 1 \geq r$ for any $r \geq 0$, then we can tell

whether $OPT_k(X) \leq \frac{|X|\sigma^2(X)}{(r+1)}$. We can find $|X|\sigma^2(X)$ in polynomial time. Also, by definition of $OPT_k(X)$, $OPT_k(X) \leq |X|\sigma^2(X)$. Thus, by setting r , we can find out if $OPT_k(X) \leq v$ for any $v \geq 0$. However, this problem is NP-hard for $k \geq 2$. \square

2.4.3 Computational Complexity of Worst Pair Ratio

Recall that a data set X is WPR well-clusterable for k whenever $WPR_k(X) > 1$, and WPR poorly-clusterable for k otherwise.

Theorem 3. *Given an integer k and a data set X that is WPR well-clusterable for k , we can find an optimal k -means clustering in polynomial-time.*

Proof. Since X is WPR well-clusterable, there exists a clustering C with optimal k -means loss where $split_C(X) > width_C(X)$. Thus, if two points are in different clusters of C the distance between them is at least $split_C(X)$. If they are in the same cluster, the distance between them is at most $width_C(X)$. Therefore, two points belong to the same cluster if and only if the distance between the points is less than $split_C(X)$. Given the value of $split_C(X)$, we go through all pairs of points and determine which belong to the same cluster. Thus, we find clustering C .

Since $split_C(X)$ is a distance between some pair of points in X , it is one of at most $\binom{n}{2}$ values corresponding to all possible pairwise distances between data points, where n is the number of points in X . We try to find a clustering using each potential value of $split_C(X)$.

The following algorithm is used:

Algorithm 2. *Clustering WPR well-clusterable data*

INPUT: A data set X and $k \geq 2$ such that X is WPR well-clusterable for k .

OUTPUT: An optimal k -means clustering of X .

Let P be the set of pairwise distances between points in X

for each $split \in P$

 find a clustering C' of X by the following procedure:

```

for each pair of points in  $\{a, b\} \subseteq X$ 
  if  $\|a - b\| < split$ 
     $a$  and  $b$  belong to the same cluster in  $X$ 
find  $width_{C'}(X)$  and  $split_{C'}(X)$ 
if  $width_{C'}(X) < split_{C'}(X)$  and  $C'$  has  $k$  clusters
  return  $C'$ 

```

Note that only one of the clusterings for which $width_{C'}(X) < split_{C'}(X)$ is going to have k clusters. This is because using this scheme, as the value of $split$ increases clusters can merge, but not lose points. If X is WPR well-clusterable for k , then there exists an optimal k -means clustering where the split is greater than the width. When $split = split_k(X)$, that clustering is found, since there is a unique clustering with the split greater than the width for fixed split. Note that to speed up the procedure we can sort the elements in P and binary search through the set until the desired number of clusters is reached.

Algorithm 2 shows that when a data set is WPR well-clusterable for k , then there is a unique k -means optimal clustering C that maximizes the split over width ratio. To simplify future discussion of WPR well-clusterable data sets, we let $split_k(X) = split_C(X)$ and $width_k(X) = width_C(X)$, when X is WPR well-clusterable for k .

Since there are at most $\binom{n}{2}$ potential values of $split$ and finding a clustering and its width given a value of $split$ is done in $O(\binom{n}{2})$, the running time of Algorithm 2 is $O(n^4)$. \square

Algorithm 2 can be used to find an appropriate number of clusters for a given data set, given that the data set is WPR well-clusterable for some $k \geq 2$. For each potential value of $split$, we can find the corresponding clustering C , and check wherever $split_C(X) > width_C(X)$, if so, then the number of clusters in C is a good choice for the number of clusters into which X should be separated. Binary search on the potential values of $split$ can be used to speed up the procedure. We can generalize this idea, and say that the number of clusters

in clustering C is good if $WPR_C(X) \geq t$, for some threshold t .

We show that it is unlikely that there is a polynomial-time algorithm for finding the optimal k -means clustering for a data set that is WPR poorly-clusterable for k .

Lemma 2. *It is an NP-hard problem to find an optimal k -means clustering given a data set which is WPR poorly-clusterable for k , where $k \geq 2$.*

Proof. Finding the optimal k clustering for $k \geq 2$ is NP-complete[4]. Suppose that there is a polynomial-time algorithm that finds an optimal k -means clustering of X for $k \geq 2$ whenever X is WPR poorly-clusterable for k . By Theorem 3, there is a polynomial-time algorithm that finds the optimal k -means clustering whenever X is WPR well-clusterable for k . Running the two algorithms in parallel, terminating when the first one of the them terminates, yields a polynomial-time algorithm for finding an optimal k -means clustering for $k \geq 2$. Therefore, it is NP-hard to find the optimal k -clustering of X whenever X is WPR-poorly-clusterable. \square

If we are given the value of x and y , can we then find an optimal clustering in polynomial-time? We cannot, unless $P=NP$.

Theorem 4. *Unless $P=NP$, there is no polynomial-time algorithm for finding the optimal k -means clustering of a data set X given the width and split of one of the optimal k -means clusterings, for $k \geq 2$.*

Proof. Assume that, given the correct value of the width and split of some k -means optimal clustering C , an optimal k -means clustering of X can be found in time n^j for some $j \geq 1$ using some Turing machine M . Then there exists a Turing machine T that takes as input a data set and simulates M for up to n^j steps for each potential pair of values for split and width. It then selects the clustering with the minimal loss. Since there are $\binom{n}{2} \cdot \binom{n}{2}$ potential values for pairs of width and split, T is a Turing machine that finds an optimal k -

means clustering in polynomial-time. Since clustering for k -means with $k \geq 2$ is NP-hard[4], clustering is NP-hard even given the split and width of some optimal k -means clustering. \square

We look at the problem of finding an upper bound on the width of an optimal clustering. Given data set X , let $w = \min\{width_C(X) \mid C \text{ a clustering of } X\}$. Finding w is NP-hard for $k \geq 3$ ([1],[6]). In [6], it is pointed out that for $k = 2$ the problem can be reduce to determining whether or not a graph is bipartite. We explain how. Start with an empty graph G . For each point in X , add a vertex to G . If the distance between two points is greater than a given threshold d , then place an edge between the associated vertices. Then there exists a bipartition of G if and only if there is a 2 clustering of X where the distance between any two points within a cluster is at most d .

We now look at the problem of finding $s = \max\{split_C(X) \mid C \text{ a clustering of } X\}$. As shown by Narashinhan et. al. finding s can be solved in polynomial-time [11]. The same result is also found by Asano et al. [1].

2.5 Comparisons of Notions of Clusterability

We perform a pairwise analysis of the notions of clusterability. Let A and B be notions of clusterability. Good clusterability by notion A does not imply good clusterability by notion B if there exist data sets with arbitrarily good A clusterability but arbitrarily bad B clusterability. Otherwise, good clusterability by notion A implies good clusterability by notion B . Notions A and B are equivalent if good clusterability by notion A implies good clusterability by notion B and good clusterability by notion B implies good clusterability by notion A . For all presented notions of clusterability, we found that no two notions are equivalent. In addition, many interesting one-directional implications were discovered.

2.5.1 Separability versus Variance Ratio

In this section, we explore the relationship between separability and variance ratio. First, we show that for $k \geq 3$, good VR clusterability does not entail good separability clusterability. Then, we will show that for $k = 2$ the two notions are equivalent, and characterize the relationship between these notions for $k \geq 3$. We use the latter result to prove that good separability clusterability implies good VR clusterability.

Theorem 5. *Given $z > 0$, $0 \leq \epsilon < 1$, and $k \geq 3$, there is a data set X such that $VR_k(X) \geq z$ and $S_k(X) \geq \epsilon$.*

Proof. Arrange a data set so that $S_{k-1}(X) \geq \epsilon$. Then add another point sufficiently far away to increase the between-cluster variance so that $VR_k(X)$ is at least z . Place the last added point sufficiently far so that it has its own cluster in any optimal k -means and any optimal $(k - 1)$ -means clustering. Therefore, the remaining points have the same clustering as in an optimal $(k - 1)$ -means clustering. The singleton cluster does not effect the k -means loss or the $(k - 1)$ -means loss. Therefore, $S_k(X) \geq \epsilon$. \square

We now present a complete characterization of the relationship between separability and VR. We show that the notions are equivalent for $k = 2$ and present the more complex relationship for $k \geq 3$.

Theorem 6. *For any data set X , $VR_2(X) = \frac{1-S_2(X)}{S_2(X)}$.*

Proof. We have the following:

$$\sigma^2(X) = W_2(X) + B_2(X)$$

$$W_2(X) = \frac{OPT_2(X)}{|X|} = \frac{S_2(X)\sigma^2(X)|X|}{|X|} = S_2(X)\sigma^2(X)$$

Thus,

$$\begin{aligned}
VR_2(X) &= \frac{B_2(X)}{W_2(X)} \\
&= \frac{\sigma^2(X) - W_2(X)}{W_2(X)} \\
&= \frac{\sigma^2(X) - S_2(X)\sigma^2(X)}{S_2(X)\sigma^2(X)} \\
&= \frac{1 - S_2(X)}{S_2(X)}
\end{aligned}$$

□

Therefore, as $S_2(X)$ decreases, $VR_2(X)$ increases. Since $S_2(X) = \frac{1}{VR_2(X)+1}$, as $VR_2(X)$ increases, $S_2(X)$ decreases. So for two clusters, a data set X has good separability if and only if it has good VR.

Now, consider the behavior for $k = 3$. $W_3(X) = \frac{OPT_3(X)}{|X|}$. $S_3(X) = \frac{OPT_3(X)}{OPT_2(X)}$, so $S_3(X) = \frac{OPT_3(X)}{S_2(X)|X|\sigma^2(X)}$. Thus, $OPT_3(X) = S_2(X)S_3(X)|X|\sigma^2(X)$, giving $W_3(X) = S_2(X)S_3(X)\sigma^2(X)$.

$$\begin{aligned}
VR_3(X) &= \frac{B_3(X)}{W_3(X)} \\
&= \frac{\sigma^2(X) - W_3(X)}{W_3(X)} \\
&= \frac{\sigma^2(X) - S_2(X)S_3(X)\sigma^2(X)}{S_2(X)S_3(X)\sigma^2(X)} \\
&= \frac{1 - S_2(X)S_3(X)}{S_2(X)S_3(X)} \\
&= \frac{1 - \frac{1}{VR_2(X)+1}S_3(X)}{\frac{1}{VR_2(X)+1}S_3(X)} \\
&= \frac{VR_2(X) + 1}{S_3(X)} - 1
\end{aligned}$$

Therefore, we can express $VR_3(X) = \frac{1}{S_2(X)S_3(X)} - 1$, only in terms of $S_2(X)$ and $S_3(X)$, or as $VR_3(X) = \frac{VR_2(X)+1}{S_3(X)} - 1$, in terms of $VR_2(X)$ and $S_3(X)$. Similarly, we can express $S_3(X) = \frac{VR_2(X)+1}{VR_3(X)+1}$, or as $S_3(X) = \frac{1}{S_2(X)(VR_3(X)+1)}$.

Lemma 3. For $k \geq 3$, $W_k(X) = S_2(X)S_3(X) \cdots S_k(X)\sigma^2(X)$.

Proof. By the above, the lemma hold for $k = 3$. Assume it holds for all $k = j - 1$. Then $W_{j-1}(X) = S_2(X)S_3(X) \cdots S_{j-1}(X)\sigma^2(X)$. $OPT_j(X) = |X|W_j(X)$, so $W_j(X) = \frac{1}{|X|}OPT_j(X)$. Now, $S_j(X) = \frac{OPT_j(X)}{OPT_{j-1}(X)}$, so $OPT_j(X) = S_j(X)OPT_{j-1}(X)$. Therefore,

$$\begin{aligned}
W_j(X) &= \frac{1}{|X|} S_j(X) OPT_{j-1}(X) \\
&= \frac{1}{|X|} S_j(X) S_{j-1}(X) OPT_{j-2}(X) \\
&= \frac{1}{|X|} S_j(X) S_{j-1}(X) S_{j-2}(X) OPT_{j-3}(X) \\
&\quad \vdots \\
&= \frac{1}{|X|} S_j(X) S_{j-1}(X) S_{j-2}(X) \cdots S_2(X) |X| \sigma^2(X) \\
&= S_j(X) S_{j-1}(X) S_{j-2}(X) \cdots S_2(X) \sigma^2(X)
\end{aligned}$$

□

Theorem 7. For $k \geq 3$, $VR_k(X) = \frac{1}{S_2(X)S_3(X)\cdots S_k(X)} - 1$.

Proof. By Lemma 1, $W_k(X) = S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X)$.

$$\begin{aligned}
VR_k(X) &= \frac{B_k(X)}{W_k(X)} \\
&= \frac{\sigma^2(X) - W_k(X)}{W_k(X)} \\
&= \frac{\sigma^2(X) - S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X)}{S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X)} \\
&= \frac{1 - S_2(X)S_3(X)\cdots S_k(X)}{S_2(X)S_3(X)\cdots S_k(X)} \\
&= \frac{1}{S_2(X)S_3(X)\cdots S_k(X)} - 1
\end{aligned}$$

□

By Theorem 7, we get that for $k \geq 3$ the following hold:

$$VR_k(X) = \frac{1}{S_2(X)S_3(X) \cdots S_k(X)} - 1$$

$$VR_k(X) = \frac{VR_{k-1}(X) + 1}{S_k(X)} - 1$$

$$S_k(X) = \frac{1}{S_2(X) \cdots S_{k-1}(X)(VR_k(X) + 1)}$$

$$S_k(X) = \frac{VR_{k-1}(X) + 1}{VR_k(X) + 1}$$

We now show that good separability clusterability implies good VR clusterability.

Theorem 8. $VR_k(X) \geq \frac{1}{S_k(X)} - 1$ for $k \geq 2$.

Proof. For $k \geq 3$, since $VR_k(X) = \frac{VR_{k-1}(X)+1}{S_k(X)} - 1$ and $VR_{k-1}(X) \geq 0$, $VR_k(X) \geq \frac{1}{S_k(X)} - 1$.

For $k = 2$, the result follows from Theorem 6. □

Therefore, good $S_k(X)$ clusterability provides a lower bound for VR clusterability.

2.5.2 Worst Pair Ratio versus Variance Ratio

We will show that it is possible to have arbitrarily high values of $VR_k(X)$ for arbitrarily low values of $WPR_k(X)$. We will then show that WPR clusterability provides a lower bound for VR clusterability.

Theorem 9. For any $x, y \geq 0$, $k \geq 2$, there exists a data set X such that $VR_k(X) \geq y$ and $WPR_k(X) \leq x$.

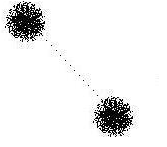


Figure 2.1: An example of a data set with good VR clusterability and poor WPR clusterability.

Proof. We first describe an example for $k \geq 3$. Consider a set of many small clusters all but two of which are far apart from each other, so that the between-cluster variance is high and the within-cluster variance is low. Exactly two of the clusters, A and B , are very close to each other, and some other cluster is sufficiently large, so that $WPR_k(X) = x$. However, by moving all pairs of clusters, except for A and B , further away from each other we increase $VR_k(X)$ arbitrarily without increasing $WPR_k(X)$. For $k = 2$, consider the example in Figure 2.1. By increasing the radius of the circles, we can make WPR arbitrarily low. Moving the circles further away from each other makes VR arbitrarily high, since most within cluster pairs are much closer to each other than most between cluster pairs. To compensate for the larger number of points on the line as the dense circles move further away from each other, we can increase the density of the circles. \square

We prove that WPR clusterability can be used to find a lower bound to VR Clusterability. First, we present an alternative formula for between-cluster variance.

Lemma 4. *Let $C = \{X_1, X_2, \dots, X_k\}$ be an optimal k -means clustering of X , where c_i is the center of mass of X_i . Then $B_k(X) = \frac{1}{|X|^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2$.*

Proof. The between-cluster variance of data set X is defined as $\sum_{i=1}^k \frac{|X_i|}{|X|} \|c_i - c\|^2$, where c is the center of mass of X . For a set $P \subseteq S$ where S is a normed vector space, $\sum_{a, b \in P} \|a - b\|^2 = |P| \sum_{a \in P} \|a - p\|^2$, where p is the center of mass of P . The between-cluster variance of X is the same as the between-cluster variance of a data set \bar{X} , having exactly $|X_i|$ points at position c_i for all $i \in \{1, 2, \dots, k\}$. The between-cluster variance of \bar{X} is $\sum_{i=1}^k \frac{|X_i|}{|X|} \|c_i - c\|^2 =$

$\frac{1}{|\bar{X}|} \sum_{x \in \bar{X}} \|x - c\|^2 = \frac{1}{|\bar{X}|^2} \sum_{a, b \in \bar{X}} \|a - b\|^2 = \frac{1}{|\bar{X}|^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2$. Since the between-cluster variance of \bar{X} is the same as the between-cluster variance of X , the result holds. \square

Theorem 10. *If X is WPR well-clusterable for k , that is, $WPR_k(X) > 1$, then $VR_k(X) > \frac{n-1}{2n}(WPR_k(X))^2$ where $|X| = n$.*

Proof. First, we show that the between-cluster variability is at least $\frac{n-1}{2n} split_k(X)^2$.

$$\begin{aligned}
B_k(X) &= \frac{1}{n^2} \sum_{i \neq j} |X_i| |X_j| \|c_i - c_j\|^2 && \text{By Lemma 4} \\
&\geq \frac{1}{n^2} \sum_{i \neq j} \|c_i - c_j\|^2 \\
&\geq \frac{1}{n^2} \sum_{i \neq j} split_k(X)^2 \\
&= \frac{1}{n^2} \binom{n}{2} split_k(X)^2 \\
&= \frac{n-1}{2n} split_k(X)^2
\end{aligned}$$

The third line holds since the distance between the centers of two clusters is at least the minimal distance between a point in one of the clusters and a point in the other cluster.

For within-cluster variability, $W_k(X) = \frac{1}{n} \sum_{i=1}^k \sum_{a \in X_i} \|a - c_i\|^2 \leq \frac{1}{n} \sum_{i=1}^k \sum_{a \in X_i} width_k(X)^2 = width_k(X)^2$. Therefore, $VR_k(X) = \frac{B_k(X)}{W_k(X)} \geq \frac{n-1}{2n}(WPR_k(X))^2$. \square

Therefore, if X is WPR well-clusterable for k , then $VR_k(X)$ is bounded from below by $\approx \frac{1}{2}(WPR_k(X))^2$. This illustrates that for large WPR clusterability, VR clusterability is high.

2.5.3 Worst Pair Ratio versus Separability

We now prove that good separability clusterability does not imply good WPR clusterability.

We then show that good WPR clusterability implies good separability clusterability.

Theorem 11. *For any $x \geq 0$, $0 < \epsilon < 1$, $k \geq 2$, there exists a data set X such that $S_k(X) \leq \epsilon$ and $WPR_k(X) \leq x$.*

Proof. By Lemma 4, there exist data sets with arbitrarily low WRP and arbitrarily high VR for $k = 2$. By Theorem 6, VR and separability are equivalent for $k = 2$. Therefore, there is a data set with arbitrarily good separability and arbitrarily poor WPR. To generalize the example for $k \geq 3$, add $k - 2$ points sufficiently far away from the remaining points and from each other so that they make distinct cluster in the optimal k -means clustering. \square

On the other hand, consider the effect of a high value of $WPR_k(X)$ on separability clusterability. High $WPR_k(X)$ means that we have k small clusters far away from each other. What happens if we have to cluster this data into $k - 1$ clusters? If $WPR_k(X)$ is sufficiently high, two of the clusters from the k clustering are joined into one. This illustrates that WPR can be used to provide a weak bound on separability clusterability.

2.5.4 Perturbation Loss versus Worst Pair Ratio

We show that if a data set X is WPR well-clusterable, then X has good PL clusterability.

First, we demonstrate that that if $WPR_k(X) > 1$ and Algorithm 1 is called on l , k , and X , such that $\frac{R}{\sqrt{l}} < \frac{\text{split}_k(X) - \text{width}_k(X)}{2}$, where R is the radius of X , the clustering returned by Algorithm 1 is an optimal k -means clustering. Next, we will show that good PL clusterability does not imply good WPR clusterability.

Lemma 5. *Given data set X and integer $k \geq 1$, if X is WPR well-clusterable for k and there exists a clustering C' that is r -close to an optimal clustering C of X , where $r < \frac{\text{split}_k(X) - \text{width}_k(X)}{2}$, then following a Lloyd step on C' , $C' = C$.*

Proof. Let p be some point in X . Let c_t be the closest center to p in C . Let c' be the closest center to c_t in C' . By the definition of a Lloyd step, it is sufficient to prove that p is closer to c' than any other center in C . Let $d \leq \text{width}_k(X)$ be the distance between point p and c_t . Let $c_{t'}$ be any other center in C . Then the distance between points p and $c_{t'}$ is $d' \geq \text{split}_k(X)$. Then the distance between p and c' is $< d + \frac{\text{split}_k(X) - \text{width}_k(X)}{2} \leq \text{width}_k(X) + \frac{\text{split}_k(X) - \text{width}_k(X)}{2} = \frac{\text{width}_k(X) + \text{split}_k(X)}{2}$. The distance between p and the center in C' closest to $c_{t'}$ is $\geq d' - \frac{\text{split}_k(X) - \text{width}_k(X)}{2} \geq \text{split}_k(X) - \frac{\text{split}_k(X) - \text{width}_k(X)}{2} = \frac{\text{split}_k(X) + \text{width}_k(X)}{2}$. Therefore, p is assigned to the cluster with center c' in C . Therefore, following a single Lloyd step a set of optimal centers is found. \square

Corollary 1. *If X is WPR well-clusterable for k , then X is $(\frac{\text{split}_k(X) - \text{width}_k(X)}{2}, 1)$ -PL clusterable.*

Note that the clustering returned by Algorithm 1 is $\frac{\text{split}_k(X) - \text{width}_k(X)}{2}$ -close to an optimal clustering whenever $l \geq \frac{4R^2}{(\text{split}_k(X) - \text{width}_k(X))^2}$. The above corollary shows that as the WPR clusterability of a WPR well-clusterable set improves, PL clusterability improves as well. On the other hand, there are data sets with low WPR clusterability but good PL clusterability.

Lemma 6. *For any $\mu > 0$, $k \geq 3$, there is a data set X with $(\mu, 1)$ -PL clusterability and WPR arbitrarily low.*

Proof. Create k groups of data. Arrange $k - 1$ of them such that each group forms a circle of radius μ and the minimal distance between two points in different groups is at least 3μ and there are at least 2 groups where the minimal separation is exactly 3μ . Then, make the last group span a circle of radius $q\mu$ for any q and place it sufficiently far apart from the other groups as well as adjust the density of all groups such that each groups makes a distinct cluster in the optimal clustering. See Figure 2.2 for an example. Then the data is $(\mu, 1)$ -PL clusterable and has arbitrarily low WPR clusterability. \square

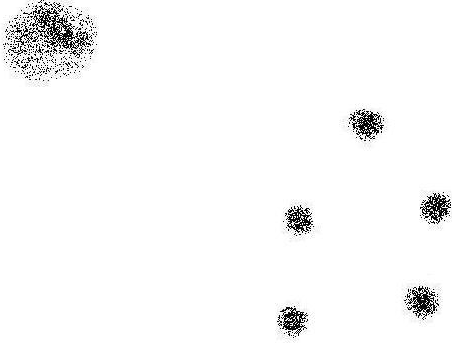


Figure 2.2: An example of a data set with good PL clusterability that is poorly-clusterable according to WPR clusterability. $k=6$.

2.5.5 Perturbation Loss versus Variance Ratio

Assume that VR clusterability is high. Does this imply that the data is well-clusterable by PL clusterability? We show that for arbitrarily high VR clusterability, we can have low PL clusterability. We then show that good PL clusterability does not imply good VR clusterability.

The following result is contained in Theorem 13, but presents an alternative solution for the $k = 3$ case. This result is simpler than the general case for arbitrary $k \geq 3$.

Theorem 12. *Given $\mu > 0$, $\alpha > 0$, and $\omega > 0$, there is a data set X with $VR_3(X) \geq \alpha$ for which there is a clustering μ -close to the optimal clustering that has loss at least $\omega k\text{-means}(C)$, where C is an optimal 3-means clustering.*

Proof. We construct data set X . See Figure 2.3 for an example. Let $n-1$ points be separated into two groups A and B of equal size such that B is A shifted to the right. The minimum distance between two points in different groups is μ . The maximum distance between two points in the same cluster is set to any arbitrarily small $\epsilon \ll \mu$. Arrange the points within the groups so that the average contribution of a point to the loss function is $(\frac{\epsilon}{2})^2$. Let a and b be the centers of mass of the two groups. Move center a and center b μ units to the

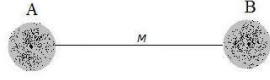


Figure 2.3: The center of cluster A is a and the center of cluster B is b .

left, as in Figure 2.4. Then all the points in these two clusters would be assigned to center b . After a Lloyd step, the center of the cluster containing the $n - 1$ points is in position c in the middle of groups A and B , and therefore the contribution of each point to the loss increases by at least $\frac{\mu^2 - 2\mu\epsilon}{4}$. Let \bar{C} be the clustering using the moved centers. Then $k\text{-means}(\bar{C}) \geq k\text{-means}(C) + \frac{1}{4}(\mu - \epsilon)^2(n - 1)$. By introducing another point far from the two clusters, we can make $B_3(X)$ arbitrarily large without effecting $W_3(X)$. Therefore, we can get arbitrarily large values of $\frac{B_3(X)}{W_3(X)}$. Since the average contribution of a point in A or B is $\frac{\epsilon}{2}$, $k\text{-means}(C) = \frac{\epsilon^2}{4}(n - 1)$.

$$\begin{aligned}
 k\text{-means}(\bar{C}) &\geq k\text{-means}(C) + \frac{1}{4}(\mu^2 - 2\mu\epsilon)(n - 1) \\
 &= \frac{\epsilon^2}{4}(n - 1) + \frac{1}{4}(\mu^2 - 2\mu\epsilon)(n - 1) \\
 &= \frac{\epsilon^2}{4}(n - 1) + \frac{\epsilon^2}{4} \cdot \frac{1}{4}(\mu^2 - 2\mu\epsilon) \frac{4}{\epsilon^2}(n - 1) \\
 &= \left(1 + \frac{\mu^2 - 2\mu\epsilon}{\epsilon^2}\right) k\text{-means}(C)
 \end{aligned}$$

As ϵ goes to 0, $(1 + \frac{\mu^2 - 2\mu\epsilon}{\epsilon^2})$ goes to infinity. In addition, the between-cluster variance does not change and the within-cluster variance decreases, thus the VR clusterability improves. Therefore, there is an ϵ such that $k\text{-means}(\bar{C}) \geq \omega k\text{-means}(C)$ and VR clusterability is at least α . □

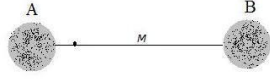


Figure 2.4: a and b moved by μ units left.

We now prove the result of $k \geq 3$.

Theorem 13. *Given $\mu > 0$, $\alpha > 0$, $\omega > 0$, and an integer $k \geq 3$, there is a data set X with $VR_k(X) \geq \alpha$ having n elements for which there is a clustering μ -close to the optimal clustering that has loss at least ωk -means(C) where C is an optimal k -means clustering.*

Proof. We construct X . In this example, $n - 1$ points in X are arranged into $k - 2$ clusters around a central cluster as in Figure 2.5. The total number of points in these clusters is $n - 1$ and each of the $k - 1$ clusters has an equal number of points. If $k - 1$ does not divide n then the clusters will have only approximately the same size. This makes little difference when n is large enough. Select ϵ and δ where $\delta < \epsilon \ll \mu$. Make the width of the clusters δ . Arrange the points within the $k - 1$ clusters so that the average contribution of a point in each of them is $(\frac{\delta}{2})^2$. The distance between a center of an outer cluster and the center of the middle cluster is $\mu - \epsilon$. Move the centers of the outer clusters by μ as in Figure 2.6. Call the clustering with the moved centers \hat{C} . Then the average increase in contribution of a point in an outer cluster to the loss function is greater than $\mu^2 - 2\mu\epsilon$. Therefore, k -means(\hat{C}) $\geq k$ -means(C) + $\frac{k-2}{k-1}(\mu^2 - 2\mu\epsilon)(n - 1)$. Place one point very far away, and it becomes its own cluster and increases $\frac{B_k(X)}{W_k(X)}$ arbitrarily. In particular, $\frac{B_k(X)}{W_k(X)}$ can be increased up to α . Since the average contribution of a point in the large $k - 1$ clusters is $\frac{\delta^2}{4}$, k -means(C) = $\frac{\delta^2}{4}(n - 1)$.

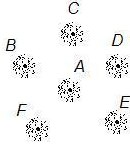


Figure 2.5: An example of a data set with high VR clusterability but low PL clusterability.

$$\begin{aligned}
 k\text{-means}(\hat{C}) &\geq k\text{-means}(C) + \frac{k-2}{k-1}(\mu^2 - 2\mu\epsilon)(n-1) \\
 &= \frac{\delta^2}{4}(n-1) + \frac{k-2}{k-1}(\mu^2 - 2\mu\epsilon)(n-1) \\
 &= \frac{\delta^2}{4}(n-1) + \frac{4(k-2)}{\delta^2(k-1)} \frac{\delta^2}{4}(\mu^2 - 2\mu\epsilon)(n-1) \\
 &= \left(1 + \frac{4(k-2)}{\delta(k-1)}\right)(\mu^2 - 2\mu\epsilon)k\text{-means}(C)
 \end{aligned}$$

Note that ϵ can be arbitrarily small. As δ goes to 0, $1 + \frac{4(k-2)}{\delta^2(k-1)}(\mu^2 - 2\mu\epsilon)$ goes to infinity. Therefore, we can choose δ so that $k\text{-means}(\hat{C}) \geq \omega k\text{-means}(C)$. When δ decreases, the between cluster variance does not change and the within cluster variance decreases, so VR clusterability improves. \square

A data set has very good PL clusterability if it is (μ, f) -PL clusterable for large μ and slow-growing function f . In particular, if $f(\eta) = 1$ for all $\eta \leq \mu$ for some large μ then the data set is well-clusterable by PL clusterable. We show that when this occurs we cannot claim that VR clusterability is high.

Theorem 14. *For any $k \geq 2$ and $\mu > 0$, there exists a data set that is $(\mu, 1)$ -PL clusterable*

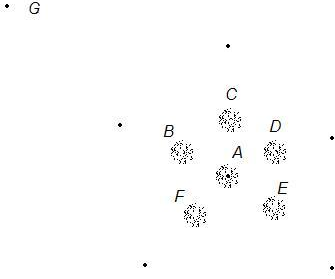


Figure 2.6: An example of a data set with high VR clusterability but low PL clusterability. The centers in the outer clusters are moved by μ units.

for k but with VR clusterability arbitrarily close to the worst possible VR for k clusters.

Proof. Consider a data set with the worst possible VR clusterability for k clusters. In an optimal k -clustering, each point is no further from its center than any other center. Perturb the points slightly so that each point is strictly closer to its center than to any other center. Since this perturbation can be arbitrarily small, it does not have much effect on the VR. Scale the data set so that the minimal difference in the distance between a point and its center and a point and its second closest center is larger than 2μ . The resulting data set is $(\mu, 1)$ -PL clusterable. Since scaling does not effect the VR (since the between and within cluster variance are multiplied by the same constant), the VR of this data set is arbitrarily close to the worst possible VR for k clusters. \square

The proof of Theorem 14 highlights a weakness of PL clusterability - its dependence on the scaling of data sets. For future work, it would be interesting to consider a variation of (μ, f) -PL clusterability where μ is scaled by some function of the distances in the data set, such as the variance of the data set. More on scale invariance appears in the next chapter.

2.5.6 Perturbation Loss versus Separability

If a data set has good PL clusterability, does that mean that it has good separability clusterability? We show that a data set can have good PL clusterability and poor separability

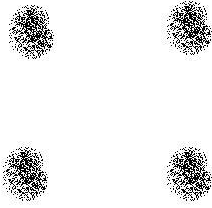


Figure 2.7: An example of data with good PL clusterability but low separability clusterability.

clusterability.

Theorem 15. *For any $\mu > 0$ and $k \geq 3$, there is a data set X with $(\mu, 1 + \frac{1}{k-1})$ -PL clusterability that and $S_k(X) \geq \frac{k-2}{k-1}$.*

Proof. Consider a set of well-separated $k - 1$ clusters all at a distance of more than μ from each other, so that moving each center by μ in any direction does not change the $(k - 1)$ -clustering. The $k - 1$ clusters are identical. See Figure 2.7. The $k - 1$ clusters are placed sufficiently far away from each other so that if we are to cluster X optimally into k clusters, then one of the $k - 1$ clusters is separated into 2. Since each of the $k - 1$ clusters are identical, they each contribute α to the loss of the $k - 1$ clustering for some $\alpha \in R$. In particular, the optimal loss of clustering with $k - 1$ clusters is $OPT_{k-1}(X) = (k - 1)\alpha$. Then the optimal loss of clustering with k clusters is $OPT_k^2(X) \geq (k - 2)\alpha$. Therefore, $\frac{OPT_k(X)}{OPT_{k-1}(X)} \geq \frac{(k-2)\alpha}{(k-1)\alpha} = \frac{k-2}{k-1}$. In the optimal k -clustering, $k - 2$ of the clusters do not change if centers are moved by at most μ units. In the remaining two clusters, points can switch centers between these clusters if centers are moved by at most μ units, which causes an increase of less than $\frac{1}{k-1}OPT_k(X)$ to the loss of the clustering. \square

As k grows, we have examples of arbitrarily poor separability clusterability for arbitrarily good PL clusterability.

We now show that good separability does not imply good PL.

Theorem 16. *Given $\mu > 0$, $\alpha > 0$, $\omega > 0$, and an integer $k \geq 2$, there is a data set X with $S_k(X) \leq \alpha$ having n elements for which there is a clustering μ -close to the optimal clustering that has loss at least ωk -means(C), where C is an optimal k -means clustering.*

Proof. Consider clusters A and B as in the proof of Theorem 12 (see Figure 2.5). By decreasing the radius of these clusters, we both improve the separability of the data set and worsen its PL clusterability. For an example on great than or equal to 3 clusters, use the data set in the proof of Theorem 13 (see Figure 2.5), ignoring the outlier cluster. As in the previous example, decreasing the radius of the clusters both improves the separability of the data set and worsens its PL clusterability. This is because when the radius is sufficiently small, the optimal $(k - 1)$ -clustering merges two of the clusters, leaving the rest unchanged. Therefore, the effect of the distance between the clusters becomes more significant as the cluster radius decreases. □

2.6 Conclusions

We have presented three notions of clusterability from the literature. We found some interesting differences on the computational hardness of these notions of clusterability. Separability and VR clusterability are computationally hard to find, while determining whether a data set is WPR well-clusterable is computationally tractable. By Ostrovsky et al. [12], data that is well-clusterable by separability clusterability is easier to cluster well. When data is WPR well-clusterable it can be clustered optimally in polynomial-time. However, by VR clusterability, according to Zhang’s [13] empirical study, it is easier to cluster data well when it is poorly-clusterable.

We introduced a new notion of clusterability, based on point perturbation, as well as variations of this notion. We found that all four of the presented notions are distinct. We have also found many one-directional implications. The following table summarizes our

comparisons. Cell (A, B) indicates whether good clusterability by measure A implies good clusterability by measure B .

	Separability	VR	WPR	PL
Separability	-	✓	x	x
VR	x	-	x	x
WPR	✓	✓	-	✓
PL	x	x	x	-

Observe that WPR is the strongest notion of clusterability - if a data set is WPR well-clusterable, then it has good clusterability by all the other notions as well. The only other positive implication is that good separability clusterability implies good VR clusterability. Therefore, WPR is the strongest notion of clusterability, followed by separability, followed by VR and PL. In many of the comparisons, we make use of the number of points or the number of clusters being arbitrarily large. It would be interesting to analyze these relationships when these parameters are fixed. Also, the hardness of PL clusterability is yet to be determined. We introduced a number of variants on PL clusterability. It would be interesting to explore their computational complexity and how they compare with other notions of clusterability.

Chapter 3

Clustering Quality

3.1 Introduction

There is a variety of clustering techniques and heuristics, which often find different clusterings of the same data set. Users often need to compare the quality of clusterings obtained by different methods. Perhaps more importantly, users need to determine whether a given clustering is sufficiently good, as it is possible that there are no good clusterings for a given data set. It is therefore surprising that there are no standard criterion for evaluating the quality of clusterings. Furthermore, to our best knowledge the concept of a clustering quality measure has never been previously formalized. In this chapter, we take the first steps to formulate a theoretical basis for clustering quality evaluation.

We present three clustering quality measures. The first measure, variance ratio, can be used for all clusterings. The second measure we introduce, called separability, is based on the clusterability notion of separability by Ostrovsky et al. [12] Separability works with loss-based clustering. We then introduce clustering quality measures for center-based clustering. We also present generalizations and alternative formalizations of the measures.

To begin the theory of clustering quality measures, we introduce a set of axioms. We

then introduce our clustering quality measures. For each clustering quality measure, we demonstrate that it satisfies the axioms of clustering quality measures. For a clustering quality measure to be useful, it is important that the quality of a clustering using the measure can be computed quickly. For each clustering quality measure we introduce, we show that the quality of a clustering using that measure can be computed in polynomial time. In addition, we also present generalizations and alternative formalizations of the measures.

In this chapter, we work in a very general setting. We assume that our set of points is $X = \{1, 2, \dots, n\}$. To define the distances between points in X , we use a distance function. A function $d : X \times X \rightarrow \mathbf{R}$ 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 $d(x_i, x_j) = d(x_j, x_i)$. Observe that $d(x, y) = 0$ if and only if $x = y$.

3.2 Previous Work

Clustering quality measures are closely related to clustering functions and clusterability. Our work on clustering quality measures makes use of ideas developed for these related concepts.

3.2.1 Clustering Functions

A clustering function is a function that takes a distance function over a data set and outputs a partition of that data set. In “An Impossibility Theorem for Clustering,” Kleinberg [7] addresses the question of whether there are any meaningful clustering functions. He proposed three axioms of clustering functions: scale invariance, richness, and consistency. He then demonstrates that no function can satisfy these three axioms simultaneously, concluding that it is not possible to axiomatize clustering functions. Since some of these axioms are relevant to our work, we present them in detail.

Scale invariance requires that the output of a clustering function be unaffected by

uniform scaling of the input.

Definition 9 (Function Scale Invariance). *Let d be a distance function. Let $d'(x, y) = \alpha d(x, y)$ for all $x, y \in X$ and some $\alpha > 0$. A function f is scale-invariant if $f(d) = f(d')$ for all such d and d' .*

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

Definition 10 (Function Richness). *A function f is rich if for each partition p of X , there exists a distance function d over X so that $f(d) = p$.*

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.

Definition 11 (Consistent Variant). *Let C be a clustering. Distance function d' is a C -consistent variant of d if $d'(x, y) \leq d(x, y)$ for $x \sim_C y$, and $d'(x, y) \geq d(x, y)$ for $x \not\sim_C y$, for all $x, y \in X$.*

Definition 12 (Function Consistency). *A function f is consistent if $f(d) = f(d')$ whenever d' is an $f(d)$ -consistent variant of d .*

Kleinberg's conclusion that the inconsistency of these axioms means that clustering functions cannot be axiomatized was premature, since we can demonstrate that consistency has some counter-intuitive consequences. In Figure 3.1, we show a good 6-clustering. On the right hand-side, we show a consistent change of this 6-clustering. Notice that resulting data has a 3-clustering, that is arguably better than the original 6-clustering. Therefore, it is undesirable to reject a function for partitioning this data set into 3 clusters.

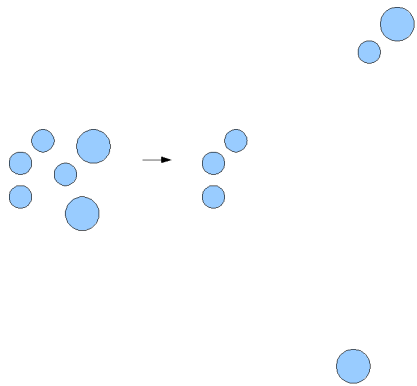


Figure 3.1: A consistent change of a 6-clustering. After the change, the quality of the original clustering decreases.

As shown by Kleinberg [7], scale invariance and richness are consistent. However, it is easy to show that these two axioms are insufficient, as we can find functions that are scale invariant and rich, but do not output meaningful clusterings. Having demonstrated that it might still be possible to axiomatize clustering functions, we do not focus on doing so in this thesis. On the other hand, we use the ideas in Kleinberg’s scale invariance and consistency axioms to develop some of the axioms of clustering quality measures.

3.2.2 Clusterability

Clusterability is another concept closely related to clustering quality. A clustering quality measure evaluates the quality of a specific clustering. A notion of clusterability, on the other hand, determines how much clustered structure there is in a data set. In Chapter 2, notions of clusterability depend on the quality of some optimal clusterings. This approach is consistent with the approach that often appears in the literature. When a measure of clusterability depends on the optimal clustering, it is often NP-hard to find the degree of clusterability (As shown in Chapter 2). Using clustering quality measures, we present an alternative approach to notions of clusterability that

gives greater flexibility. We discuss this in more detail in Section 3.6.

On the other hand, we make use of the notions of clusterability discussed in Chapter 2 to define our measures of clustering quality. We present measures of clustering quality based on the variance ratio, separability, and relative margin notions of clusterability. Notice that notions of clusterability cannot be used directly as measures of clustering quality, since measures of clustering quality evaluate the quality of specific clusterings.

3.3 Axioms of Clustering Quality Measures

A clustering quality measure is a function that is given a clustering C and distance function d , and returns a non-negative real number. The range of a clustering quality measure over non-trivial clusterings is an open interval. For a clustering quality measure to be meaningful, we need to introduce additional constraints. In this section we present four axioms of clustering quality measures. Note that in our discussions and definitions, we assume that quality measure m gives higher values to better clusterings. When the reverse is true, simply reverse the direction of inequalities that compare the quality of clusterings.

The first axiom, scale invariance, is based on Kleinberg's scale invariance axiom for clustering functions [7]. Scale invariance requires that clustering quality be unaffected by uniform scaling of all pairwise distances. This requirement is one of the main reasons why standard loss functions, such as k -means or k -median, should not be used as clustering quality measures.

Definition 13 (Scale Invariance). *Let α be a positive constant. Let d and d' be distance functions such that for all $x, y \in X$, $d'(x, y) = \alpha d(x, y)$. Let C be a clustering. A quality measure m is scale invariant if for all such d, d' , and C , $m(C, d) = m(C, d')$.*

The next axiom makes sure that quality measures are independent of point description.

This axiom captures the idea that the quality of any clustering should depend only on the pairwise distances between points.

Definition 14 (Isomorphism Invariance). *Let C and C' be clusterings such that $x \sim_C y$ if and only if $\phi(x) \sim_{C'} \phi(y)$, for ϕ a distance-preserving isomorphism from X to X . A quality measure m is isomorphism invariant if $m(C, d) = m(C', d)$ for all such C and C' .*

The next axiom is more complex than the previous ones. We motivate the next axiom by first discussing simpler variants. If we translate Kleinberg’s [7] clustering function consistency axiom to a clustering quality axiom, we get the requirement that the quality of a clustering does not worsen when pairs of points within clusters are moved closer together and pairs of points in different clusters are moved further apart.

Definition 15 (Consistency). *A quality measure m is consistent if $m(C, d) \geq m(C, d')$ for all clusterings C and distance functions d , whenever d' is a C -consistent variant of d .*

However, consistency does not make a good clustering quality axiom since it has some counter-intuitive consequences. For instance, if some clusters are moved very far away from all others, a smaller number of clusters may be more appropriate (See Figure 3.1). We do not wish to reject measures by which the quality of clustering drops following such a consistent change, since a clustering quality measure should indicate how good a clustering quality measure is in the universe of all clusterings over all data sets.

A simple modification of 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. The problem with this modification is that it 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 in a non-uniform manner.

We present a version of consistency, called local consistency, which does not have the problem of consistency discussed above and fully applies to Euclidean spaces. The intuition behind this axiom is as follows. Consider fixing a single point within each cluster. Then, move the clusters away from each other, so that the pairwise distances between all fixed points is scaled uniformly. Lastly, shrink each cluster, so that the distances between points in the same cluster are changed uniformly. The distance between points in different clusters cannot decrease; however, we do not require that these distances be scaled uniformly.

Definition 16 (Locally Consistent Variant). *Given distance functions d and d' and a clustering C , we say that d' is a C -locally consistent variant of d if*

- *For every cluster l of C there is a constant $c_l \leq 1$, such that for all $x, y \in l$, $d'(x, y) = c_l d(x, y)$.*
- *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 l , there exists a constant $c \geq 1$ such that, for every $p_l, p_{l'}$, $d'(p_l, p_{l'}) = c \cdot d(p_l, p_{l'})$.*

Definition 17 (Local Consistency). *A quality measure m is locally consistent if $m(C, d') \geq m(C, d)$ whenever d' is a C -locally consistent variant of d .*

Constant functions, as well as a function that simply returns the number of points in the data set, satisfy scale invariance, isomorphism invariance, and local consistency. We therefore introduce another axiom, called fullness, which forces the quality measure to be responsive to distances. The axiom of fullness requires that arbitrarily good (or bad) non-trivial clusterings should be obtainable by bounding the distances between pairs of points that are within clusters, and distances between pairs of points belonging to different clusters.

Definition 18 (Fullness). *Let C be a k -clustering of data set X where each cluster has at least two points and $k \geq 2$. A clustering quality measure m is full if:*

- For any $M \in \text{range}(m)$, there exist $a, b \in \mathbb{R}^+$ such that for every distance function d where for all $x \sim_C y$, $d(x, y) \leq a$ and for all $x \not\sim_C y$, $d(x, y) \geq b$, $m(C, d) > M$ (or $m(C, d) < M$, if lower values of m indicate better clustering.)
- For any $M \in \text{range}(m)$, there exist $a, b \in \mathbb{R}^+$ such that for every distance function d where for all $x \sim_C y$, $d(x, y) \geq a$ and for all $x \not\sim_C y$, $d(x, y) \leq b$, $m(C, d) < M$ (or $m(C, d) > M$, if lower values of m indicate better clustering.)

Fullness captures the idea that the quality of a clustering can improve by tightening the clusters and moving them further apart from each other. This is distinct from consistency and local consistency, where we require that under certain conditions the quality of a clustering should not get worse. By specifying how a clustering quality can be made better or worse, fullness gives some indication for what makes a clustering good or bad, without being too restrictive.

3.3.1 Axioms Non-redundancy

We now illustrate that all the axioms are non-redundant. To do that, we demonstrate functions that do not make good clustering quality functions and satisfy all but one of the axioms presented.

Non-redundancy of Scale Invariance

A *clustering loss function* is a function that takes a clustering of a data set and a distance function over the data set, and returns a real number. Many clustering algorithms attempt to find clusterings that minimal a specific loss function. Therefore, it is natural to propose to use the loss of a clustering as a measure of clustering quality. The k -means and k -median are commonly used clustering loss functions. To use these loss functions as quality measures, we let the loss of a clustering be its quality. For the k -means and

k -median loss functions, such quality measures satisfy isomorphism-invariance, local consistency, and fullness. However, clustering loss functions often depend on the scaling of the data. In particular, k -means and k -median are not scale invariant. That is, by scaling all pairwise distances uniformly, any loss function value can be obtained. Therefore, such loss functions are inappropriate for comparing clusterings over different distance functions. Since clustering quality measures should compare the quality of clusterings over different data sets in a meaningful way, clustering loss functions such as k -means and k -median do not make good clustering quality measures.

Non-redundancy of Isomorphism Invariance

To show that isomorphism invariance is non-redundant, consider any clustering quality measure m that satisfies all four axioms. We present examples of such measures in the next section. We will now create a clustering quality measure m' that satisfies scale invariance, local consistency, and fullness, but does not satisfy isomorphism invariance. Using only the descriptions of the points, m selects two clusters. For instance, it can select the two clusters with the highest maximal point values. Then, m' applies m only on these two clusters, ignoring the rest of the clusters. Due to the method of selecting the two clusters on which m is applied, any two clusters can be selected following a distance preserving isomorphism on the data set. Setting m to be any of the clustering quality measures introduced in the next section, it is easy to construct a clustering where m gives a different value when applied to each pair of clusters. Therefore, m' is not isomorphism invariant. Since m' selects the two clusters on which to apply m independently of distances between points, this measure does not make a good clustering quality measure. That is, the two clusters selected may not be representative of the features of the clustering. Note that m' is scale-invariant, locally consistent, and full.

Non-redundancy of Fullness

To see that fullness is non-redundant, notice that any constant function, or a function that returns the number of points in the data set, satisfies scale-invariance, isomorphism-invariance, and local consistency. Clearly, such functions do not make good clustering quality measures. Note that since these functions return the same value for all clustering with no regard to distances between points, these measures are not full.

Non-redundancy of Local Consistency

To illustrate that local consistency is non-redundant, consider a function with range $[0, \infty)$ that returns the ratio $\frac{\min_{y \neq C^x} d(x,y)}{\max_{y \sim C^x} d(x,y)}$ when this ratio is in the range $[0, 0.1]$ or $[1, \infty]$, and returns 100 otherwise. This is a combination of a valid quality measure with a constant function. Notice that this measure is scale invariant and isomorphism invariant. This measure is full, due to its behavior outside $(0.1, 1)$. However, given a clustering with $\frac{\min_{y \neq C^x} d(x,y)}{\max_{y \sim C^x} d(x,y)}$ in the range $[0.1, 1]$, some locally consistent variants of the clustering will have lower quality measure than the clustering.

3.4 Examples of Quality Measures Satisfying the Axioms

We now introduce quality measures that satisfy the above axioms. Each measure has specific properties that make it well suited for evaluating clusterings in different settings. Some of the measures are defined with respect to a loss function, and therefore are well-suited when it is expected that a specific loss function will capture the structure of a good clustering. We also present clustering quality measures specifically for center-based clustering.

For each quality measure, we discuss its motivation and special properties. We prove that each one satisfies the four axioms: scale invariance, isomorphism invariance, local consistency, and fullness. We also show how to find the quality of a clustering using each measure in polynomial time.

3.4.1 Variance Ratio

We begin with a general purpose clustering quality measure. This measure looks at the relationship of the between-cluster variance and the within-cluster variance. This clustering quality measure is similar to the variance ratio notion of clusterability by Zhang [13].

Standard Variance Ratio

Let d be a distance function and C a clustering. Let the *within-cluster variance* of C on d be $W(C, d) = \text{avg}_{x \sim_C y} d(x, y)$, the average distance between elements within the same cluster. Let the *between-cluster variance* of C on d be $B(C, d) = \text{avg}_{x \not\sim_C y} d(x, y)$, the average distance between elements in different clusters.

Definition 19 (Standard Variance Ratio). *The Standard Variance Ratio of C on d is*

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

Note that the range of SVR is $[0, \infty)$. Larger values of SVR indicate better clustering quality.

Axiom Satisfaction

We now show that SVR satisfies the four axioms of clustering quality measures.

Lemma 7. *Standard variance ratio is scale invariant.*

Proof. Let d and d' be distance functions such that for all $x, y \in X$, $d'(x, y) = \alpha d(x, y)$ for some constant $\alpha > 0$. Then,

$$\begin{aligned}
SVR(C, d') &= \frac{B(C, d')}{W(C, d')} \\
&= \frac{\text{avg}_{x \not\sim_C y} d'(x, y)}{\text{avg}_{x \sim_C y} d'(x, y)} \\
&= \frac{\text{avg}_{x \not\sim_C y} \alpha d(x, y)}{\text{avg}_{x \sim_C y} \alpha d(x, y)} \\
&= \frac{\alpha \cdot \text{avg}_{x \not\sim_C y} d(x, y)}{\alpha \cdot \text{avg}_{x \sim_C y} d(x, y)} \\
&= SVR(C, d)
\end{aligned}$$

□

Lemma 8. *Standard variance ratio is locally consistent.*

Proof. Let d be a distance function and C a clustering. Let d' be a C -locally consistent variant of d . Then for all $x \sim_C y$, $d'(x, y) \leq d(x, y)$ and for all $x \not\sim_C y$, $d'(x, y) \geq d(x, y)$. Therefore, $W(C, d') \leq W(C, d)$ and $B(C, d') \geq B(C, d)$. Thus, $STV(C, d') \geq STV(C, d)$. □

Lemma 9. *Standard variance ratio is full.*

Proof. The range of SVR over non-trivial clusterings is $(0, \infty)$. By setting $d(x, y) \leq 0.9$ for $x \sim_C y$ and $d(x, y) \geq b$ for all $x \not\sim_C y$, we get $SVR(C, d) > b$. Therefore, $SVR(C, d)$ can be made arbitrarily large.

By setting $d(x, y) \geq 1.1$ for all $x \sim_C y$ and $d(x, y) \leq b$ for all $x \not\sim_C y$, we get $SVR(C, d) < b$. Therefore, $SVR(C, d)$ can be made arbitrarily close to 0. □

Finally, we note that SVR is isomorphism invariant since it is independent of point description.

Time Complexity

The within-cluster variance and the between-cluster variance can be found in total time $O\binom{n}{2}$, by computing the distances between all pairs of points. Therefore, the SVR of a clustering can be computed in polynomial time.

Generalized Variance Ratio

We present a variation on standard variance ratio that works with any loss function. Let C_1 denote the 1-clustering of X . Let \mathcal{L} be any clustering loss function. Recall that a clustering loss function is a function that takes a clustering and a distance function, and returns a real number.

Definition 20 (Variance Ratio of C With Respect to \mathcal{L}). *Given a clustering C of data set X and distance function d over X , the variance ratio of C with respect to \mathcal{L} is*

$$VR_{\mathcal{L}}(C, d) = \frac{\mathcal{L}(C_1, d) - \mathcal{L}(C, d)}{\mathcal{L}(C, d)}.$$

Note that standard variance ratio is variance ratio with $\mathcal{L}(C, d) = \text{avg}_{x \sim_C y} d(x, y)$.

Loss Conformity

With loss-based clustering, desirable clusterings should have low loss. As previously noted, a first natural proposition for a quality measure of a clustering is the clustering loss function. However, many commonly used loss functions, such as k -means and k -median, depend on the scaling of the data, so they cannot be used as quality measures.

On the other hand, it is sometimes desirable that a quality measure does not contradict the 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 is not a requirement for all clustering quality measures, since often there is no relevant loss function, or the loss function is not sufficiently reliable for such a requirement. However, when the user believes that, modulo scale invariance, clusterings with lower loss are better, then loss conformity is desirable.

Whenever a quality measure satisfies this property for a clustering loss function \mathcal{L} , we say that it *conforms with \mathcal{L}* .

Definition 21 (Loss Conformity). *Let \mathcal{L} be a clustering loss function. Let C and C' be clusterings and d a distance function. A clustering quality measure m conforms with \mathcal{L} if whenever $\mathcal{L}(C, d) \leq \mathcal{L}(C', d)$, $m(C, d) \geq m(C', d)$.*

We now show that VR with respect to \mathcal{L} conforms with \mathcal{L} .

Lemma 10. *Variance ratio with respect to \mathcal{L} conforms with \mathcal{L} .*

Proof.

$$VR_{\mathcal{L}}(C, d) = \frac{\mathcal{L}(C_1, d) - \mathcal{L}(C, d)}{\mathcal{L}(C, d)}.$$

Since $\mathcal{L}(C_1, d)$ is constant over all clusterings of X , as $\mathcal{L}(C, d)$ decreases, $VR_{\mathcal{L}}(C, d)$ increases. Therefore, given two clusterings of the same data set, the clustering with lower loss has better variance ratio with respect to \mathcal{L} . □

This property allows us to view generalized variance ratio as a normalized loss function; it preserves the comparative power of the loss function on clusterings of the same distance function, and allows comparisons of clusterings over different distance functions via its scale invariance. That is, while using a loss function as a clustering quality

measure tends to be inappropriate since many are not scale invariant, using generalized variance ratio preserves the desirable quality of using a loss function as a quality measure, while being scale invariant.

3.4.2 Separability

Ostrovsky et al.[12] introduced the separability notion of clusterability, discussed in detail in Chapter 2. We wish to define a notion of clustering quality that captures similar features. Consider the $(k-1)$ -clustering C' of minimal loss that has the same clusters as a k -clustering C , except with two of the clusters in C merged. The separability of clustering C is the ratio of the loss of C over the loss of C' .

Let $C = \{C_1, C_2, \dots, C_k\}$ be some k -clustering, and \mathcal{L} a clustering loss function. For all $i \neq j$, let $C_{ij} = \{C \setminus \{C_i, C_j\} \cup \{C_i \cup C_j\}\}$ be a clustering identical to C , except with cluster C_i and C_j merged. We define separability as follows¹.

Definition 22 (Separability with respect to \mathcal{L}). *Let C be a clustering and d a distance function. Then the separability of C on d with respect to \mathcal{L} is,*

$$S(C, d) = \frac{\mathcal{L}(C, d)}{\min_{i,j} \mathcal{L}(C_{ij}, d)}.$$

Separability can be defined with respect to any loss function. As an example, loss functions of the following form can be used with separability.

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

for $\delta, \beta, \gamma \in \mathbf{R}$. For $\delta = 1, \beta = 2, \gamma = 1$, we get the k -means loss function. For $\delta = 2, \beta = 2$,

¹To define separability for clustering quality measures in a manner similar to the definition of separability as a notion of clusterability presented by Ostrovsky et al. [12], we say that C is ϵ -separable if $\frac{\mathcal{L}(C,d)}{\min \mathcal{L}(C_{ij},d)} \leq \epsilon$.

$\gamma = 1$, the loss of a cluster is its variance. For $\delta = 0$, $\beta = 1$, $\gamma = 1$, the loss of a cluster is the sum of all pairwise distances in the cluster.

Separability with k -means

Ostrovsky et al. [12] defined separability as a notion of clusterability with the k -means loss function. We discuss separability as a clustering quality measure with the k -means loss function. We prove that separability with k -means satisfies the four axioms of clustering quality measures, and discuss its various properties.

Axiom Satisfaction

We now show that separability with k -means satisfies the four axioms of clustering quality measures. Similar proofs can be used to show that separability with many other loss functions also satisfies these axioms.

Theorem 17. *Separability with k -means is scale invariant.*

Proof. Let d and d' be distance functions so that $d'(x, y) = \lambda d(x, y)$ for all $x, y \in X$, and some $\lambda \in R^+$. Therefore, $k\text{-means}(C, d') = \sum_{i=1}^k \frac{1}{|C_i|} \sum_{\{x,y\} \subseteq C_i} d'(x, y)^2 = \sum_{i=1}^k \frac{1}{|C_i|} \sum_{\{x,y\} \subseteq C_i} (\lambda d(x, y))^2 = \lambda^2 k\text{-means}(C, d)$. Therefore, for any clusterings C, C' of X , $\frac{k\text{-means}(C, d)}{k\text{-means}(C', d)} = \frac{k\text{-means}(C, d')}{k\text{-means}(C', d')}$. Therefore, it follows that $S(X, d) = S(X, d')$. \square

Theorem 18. *Separability with k -means is locally consistent.*

Proof. Let d a distance function, C a clustering, and d' a C -locally consistent variant of d . Let $C = \{C_1 \cup C_2, C_3, \dots, C_k\}$ be a clustering of X . Without loss of generality, we assume that $k\text{-means}(C_{12}) \leq k\text{-means}(C_{ij})$ for all $i \neq j$. Observe that $\sigma_d^2(Y) = \frac{1}{|Y|^2} \sum_{\{x,y\} \subseteq Y} d(x, y)^2$, is the variance of Y with respect to d , since $\sum_{\{x,y\} \subseteq Y} d(x, y)^2 = |Y| \sum_{y \in Y} d(y, \bar{y})^2$ where $\bar{y} = \frac{1}{|Y|} \sum_{y \in Y} y$. For $S \subseteq C$, we let $\sigma_d^2(S)$ denote the variance of the points in S .

Then,

$$\begin{aligned}
\frac{1}{S(C, d)} &= \frac{k\text{-means}(C_{12}, d)}{k\text{-means}(C, d)} \\
&= \frac{|C_1 \cup C_2| \sigma_d^2(C_1 \cup C_2) + |C_3| \sigma_d^2(C_3) + \dots + |C_k| \sigma_d^2(C_k)}{|C_1| \sigma_d^2(C_1) + |C_2| \sigma_d^2(C_2) + \dots + |C_k| \sigma_d^2(C_k)} \\
&= \frac{|C_1 \cup C_2| \sigma_d^2(C_1 \cup C_2) - |C_1| \sigma_d^2(C_1) - |C_2| \sigma_d^2(C_2) + \sum_{u=1}^k |C_u| \sigma_d^2(C_u)}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} \\
&= \frac{|C_1 \cup C_2| \sigma_d^2(C_1 \cup C_2) - |C_1| \sigma_d^2(C_1) - |C_2| \sigma_d^2(C_2)}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1 \\
&= \frac{\frac{1}{|C_1 \cup C_2|} \sum_{\{x, y\} \subseteq C_1 \cup C_2} d(x, y)^2 - |C_1| \sigma_d^2(C_1) - |C_2| \sigma_d^2(C_2)}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1
\end{aligned}$$

Let $r = \frac{1}{|C_1| + |C_2|}$. Then,

$$\frac{1}{S(C, d)} = \frac{r \left(\sum_{x \sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d'(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d'(x, y)^2 \right) - |C_1| \sigma_{d'}^2(C_1) - |C_2| \sigma_{d'}^2(C_2)}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1.$$

Since $|C_i| \sigma_{d'}^2(C_i) = \frac{1}{|C_i|} \sum_{\{x, y\}} d(x, y)$,

$$\frac{1}{S(C, d)} = \frac{r \left(\sum_{x \sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d(x, y)^2 \right) - \frac{1}{|C_1|} \sum_{\{x, y\} \subseteq C_1} d(x, y) - \frac{1}{|C_2|} \sum_{\{x, y\} \subseteq C_2} d(x, y)}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1.$$

Therefore,

$$\frac{1}{S(C, d)} = \frac{\left(\frac{1}{|C_1|+|C_2|} - \frac{1}{|C_1|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_1} d'(x, y)^2 + \left(r - \frac{1}{|C_2|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_2} d'(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d'(x, y)^2}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1.$$

Now, let d' be a C -locally consistent variant of d . Then,

$$\begin{aligned} \frac{1}{S(C, d')} &\geq \frac{k\text{-means}(C_{12}, d')}{k\text{-means}(C, d')} \\ &= \frac{\left(r - \frac{1}{|C_1|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_1} d'(x, y)^2 + \left(r - \frac{1}{|C_2|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_2} d'(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d'(x, y)^2}{\sum_{u=1}^k |C_u| \sigma_{d'}^2(C_u)} + 1 \\ &\geq \frac{\left(r - \frac{1}{|C_1|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_1} d(x, y)^2 + \left(r - \frac{1}{|C_2|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_2} d(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d'(x, y)^2}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1 \\ &\geq \frac{\left(r - \frac{1}{|C_1|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_1} d(x, y)^2 + \left(r - \frac{1}{|C_2|}\right) \sum_{x \sim_C y, \{x, y\} \subseteq C_2} d(x, y)^2 + \sum_{x \not\sim_C y, \{x, y\} \subseteq C_1 \cup C_2} d(x, y)^2}{\sum_{u=1}^k |C_u| \sigma_d^2(C_u)} + 1 \\ &= \frac{1}{S(C, d)} \end{aligned}$$

Line (3) follows since $\frac{1}{|C_1|+|C_2|} \leq \frac{1}{|C_2|}$, $\frac{1}{|C_1|+|C_2|} \leq \frac{1}{|C_1|}$, and $d'(x, y) \leq d(x, y)$ whenever $x \sim_C y$. Line (4) follows since $d'(x, y) \geq d(x, y)$ whenever $x \not\sim_C y$. \square

Theorem 19. *Separability with k -means is full.*

Proof. The range of separability over non-trivial clusterings is $(0, 1)$. Consider a clustering C where $d(x, y) \leq 1$ for all $x \sim_C y$ and $d(x, y) \geq b$ for all $x \not\sim_C y$. The within cluster distances bound the k -means loss of C . On the other hand, by increasing b , we can make the loss of any clustering that merges two clusters in C arbitrarily high. Therefore, $S(C, d)$ can be made arbitrarily close to 0.

To get values of $S(C, d)$ arbitrarily close to 1, set $d(x, y) \geq 1$ for all $x \sim_C y$ and $d(x, y) \leq b$ for all $x \not\sim_C y$. Then as b decreases, the ratio of the k -means loss of C over the loss of the minimal loss clustering that merges two clusters in C approaches 1. \square

Finally, we note that separability with k -means is isomorphism-invariant since it is independent of point description.

Loss Conformity

While it may seem that separability is another way to normalize loss, it should be noted that separability with k -means does not conform with the k -means loss function. The reason for this is that the k -means loss function is independent of between-cluster distances, while separability is not.

Consider the two clusterings as illustrated in Figure 3.2. The data set consists of 4 points in 1d, where the leftmost pair has distance 0.8, and the remaining distances between consecutive points are 1 unit each. The clusterings C and C^* are marked in non-dashed ovals, whereas the dashed ovals illustrated the pairing up of clusters that gives the best 2-clustering over all possible clusterings that consist of joining some pair of clusters in each of C and C^* . It is immediate that the 3-means loss of C^* is lower than the 3-means loss of C . However, since we can obtain a better 2-clustering from C^* than from C , giving less of an improvement from the 3 to the 2-clustering in C^* , C has better separability. In particular, we can show that the separability of C is approximately 0.3086 whereas the separability of C^* is no better than 0.39.

Time Complexity

Assuming that the loss function \mathcal{L} of a k -clustering with respect to d with $|X| = n$ can be evaluated in $g(n, k)$ operations where $g(n, k)$ is a polynomial function, we can

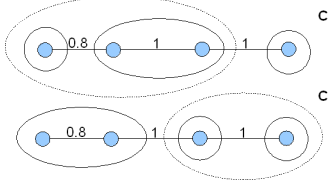


Figure 3.2: An example of two 3-clusterings, C and C^* , marked in non-dashed ovals, where the 3-means loss of C^* is smaller than the 3-means loss of C , but C has better separability than C^* .

find $S(C, d)$ in polynomial time. Since $|\{C_{ij} \mid i \neq j\}| = \binom{k}{2}$, and we compute the loss of each clustering C_{ij} in $g(n, k - 1)$ operations, we can find $S(C, d)$ in $\binom{k}{2}g(n, k - 1) + g(n, k)$ operations. If \mathcal{L} is the k -means loss function, which can be evaluated in $O(n^2)$ operations, then we can find $S(C, d)$ in $O(k^2n^2)$ operations.

Alternative Formulations

Separability with loss functions such as k -means or 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 standard variance ratio behaves differently on this aspect, since standard 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.

We can consider alternative formulations of separability which behave differently in this regard. For instance, we can define separability as $\frac{\mathcal{L}(C, d)}{\max\{\mathcal{L}(C_{ij}, d), i \neq j\}}$, which is sensitive to the maximal separation between clusters. Alternatively, we could choose to look at $\frac{\mathcal{L}(C, d)}{\text{avg}\{\mathcal{L}(C_{ij}, d) \mid i \neq j\}}$. Other variations we may consider is merging more than two clusters of C , followed by finding the minimal, maximal, or average such clustering, and computing the ratio of the loss of the original clustering over the loss of the new clustering.

3.4.3 Margins

We now introduce quality measures for center-based clustering. Therefore, we defined a center-based clustering as follows. A clustering $C = \{C_1, C_2, \dots, C_k\}$ of data set X is *center-based* if there exist points $c_1 \in C_1, c_2 \in C_2, \dots, c_k \in C_k$ such that for all $x \in X$, if $x \in C_i$ then $d(x, c_i) < d(x, c_j)$, for all $i \neq j$. That is, C is a Voronoi partition. Note that in the general setting, where the input is a distance function, we cannot use the center of mass as the center of a cluster since it is not well-defined. To allow for different approaches for determining centers, we assume that the centers of the clustering are given. Note that a center-based clustering is fully specified by its set of centers.

Relative Margin

For each point in the data set, we consider the ratio of the distance from the point to its closest center, over the distance from the point to the second closest center. We then take the average over all ratios of non-centers. Intuitively, we can view smaller margins as the points being “more sure” to which cluster they belong. This clustering quality measure is based on the relative margin notion of clusterability introduced in Chapter 2.

Definition 23 (Relative Point Margin). *Let $C = \{C_1, C_2, \dots, C_k\}$ be a center-based clustering on distance function d , with c_i the center of cluster C_i . For point $x \in X$, the relative point margin of x in C on d is $RM_{C,d}(x) = \frac{d(x, c_i)}{d(x, c_j)}$, where c_i is the closest center to x and c_j is a second closest center to x .*

Definition 24 (Relative Margin). *The relative margin of C on d is*

$$RM_d(C) = \text{avg}_{x \in X, x \neq c_i \text{ for any } i} RM_{C,d}(x).$$

The range of relative margins is $[0, 1)$. A smaller Relative Margin (RM) indicates a better clustering.

Axiom Satisfaction

We now discuss how RM satisfies the four axioms of clustering quality measures. Since relative point margin is independent of the description of the points, RM is isomorphism invariant. We show that it is also scale-invariant, locally consistent, and full.

Lemma 11. *Relative margin is scale invariant.*

Proof. Let C be a center-based clustering on distance function d . Let d' be a distance function so that $d'(x, y) = \alpha d(x, y)$ for all $x, y \in X$ and some $\alpha \in \mathbf{R}^+$. Notice that the centers in C on d' are also valid centers in C on d , that is, C is a center-based clustering on d' . Also, for any points $x, y, z \in X$, $\frac{d(x, y)}{d(x, z)} = \frac{d'(x, y)}{d'(x, z)}$. Therefore, $RM_{d'}(C) = RM_d(C)$. \square

Lemma 12. *Relative margin is locally consistent.*

Proof. Let C be a center-based clustering of distance function d . Let d' be a C -locally consistent variant of d . Since d' is a C -locally consistent variant of d , for $x \sim_C y$, $d'(x, y) \leq d(x, y)$ and for $x \not\sim_C y$, $d'(x, y) \geq d(x, y)$. Therefore, the centers of C with respect to d are valid centers for C with respect to d' . In addition, for every point $x \in C_i$ for any $1 \leq i \leq k$, $d'(x, c_i) \leq d(x, c_i)$, and $d'(x, c_j) \geq d(x, c_j)$ for $i \neq j$. Thus, $RM_{d'}(C) \leq RM_d(C)$. \square

Lemma 13. *Relative margin is full.*

Proof. The range of RM over non-trivial clusterings is $(0, 1)$. If we set $d(x, y) \leq a$ for all $x \sim_C y$, and $d(x, y) \geq 1$ for all $x \not\sim_C y$, then $RM_d(C) \leq a$. To get RM arbitrarily

close to 1, we set $d(x, y) \geq 1 - \epsilon$ for all $x \sim_C y$, and $d(x, y) \leq 1$ for all $x \not\sim_C y$, then $RM_d(C) \geq 1 - \epsilon$. \square

Time Complexity

We can find the relative point margin of all points in $O(nk)$, where k is the number of clusters in the clustering. Finding the average over all relative point margins adds linear time. Thus, the total running time for finding the RM of a clustering is $O(nk)$.

Alternative Formulations

There are many interesting variations of relative margin. We could look at the average ratio of the distance to the closest center over the distance to the r^{th} -closest center. Alternatively, we could look at the average ratio of the closest center over the average distance to all other centers. Also, instead of taking the average of ratio, we could take the maximum over all ratios.

Additive Margin

Instead of looking at the ratio of the distance to the second closest center over the ratio of the distance to the closest center, additive margin looks at the difference of these two quantities. Additive margin is similar to the perturbation loss notion of clusterability, presented in Chapter 2.

Definition 25 (Additive Point Margin). *Let d be a distance function. Let $C = \{C_1, C_2, \dots, C_k\}$ be a center-based clustering, with c_i the center of cluster C_i . For point $x \in X$, the additive point margin of x in C on d is $AM_{C,d}(x) = d(x, c_j) - d(x, c_i)$, where c_i is the closest center to x and c_j is a second closest center to x .*

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 26 (Additive Margin). *The additive margin of C on d is*

$$AM_d(C) = \frac{\text{avg}_{x \in X} AM_{C,d}(x)}{\text{avg}_{x \sim_C y} d(x, y)}.$$

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

Axiom Satisfaction

Lemma 14. *Additive margin is scale invariant.*

Proof. Let C be a center-based clustering of distance function d . Let d' be a distance function so that $d'(x, y) = \alpha d(x, y)$ for all $x, y \in X$ and some $\alpha \in R^+$. Notice that C is a center-based clustering on d' using the same set of centers. Also, for any points $x, y, z \in X$, $d'(x, y) - d'(x, z) = \alpha(d(x, y) - d(x, z))$. Thus,

$$AM_{d'}(C) = \frac{\text{avg}_{x \in X} AM_{C,d'}(x)}{\text{avg}_{x \sim_C y} d'(x, y)} = \frac{\alpha \cdot \text{avg}_{x \in X} AM_{C,d}(x)}{\alpha \cdot \text{avg}_{x \sim_C y} d(x, y)} = AM_d(C).$$

□

Lemma 15. *Additive margin is locally consistent.*

Proof. Let C be a center-based clustering of distance function d . Let d' be a C -locally consistent variant of d . Since d' is a C -locally consistent variant of d , for $x \sim_C y$, $d'(x, y) \leq d(x, y)$ and for $x \not\sim_C y$, $d'(x, y) \geq d(x, y)$. Therefore, the centers of C on d are valid centers for C on d' . Also, a C -locally consistent change can only increase the margin of each point. Combined with the fact that $\text{avg}_{x \sim_C y} d'(x, y) \leq \text{avg}_{x \sim_C y} d(x, y)$, we get that $AM_{d'}(C) \geq AM_d(C)$. □

Lemma 16. *Additive margin is full.*

Proof. The range of AM over non-trivial clusterings is $(0, \infty)$. Set $d(x, y) > 1$ for all $x \sim_C y$, and $d(x, y) \leq 1 + \epsilon$ for all $x \not\sim_C y$ and any $\epsilon > 0$. Then $AM_d(C) < \epsilon$.

To see that arbitrarily large values of AM can be obtained by setting ranges for within and between cluster distances, notice that if $d(x, y) \leq 0.9$ for all $x \sim_C y$, and $d(x, y) \geq 1 + h$ for all $x \not\sim_C y$ and any $h > 0$, then $AM_d(C) > h$. \square

Time Complexity

As with relative point margin, we can find the additive point margin of all points in $O(nk)$. Finding the average over all additive point margins adds linear time. Finding the average within-cluster distance takes $O(n^2)$ operations. Thus, the total running time for finding the AM of a clustering is $O(n^2k)$.

Alternative Formulations

We could look at the difference between the distance to the r^{th} -closest center and the distance to closest center. We could also define additive point margin as the difference of the average distance from the point to a center and the distance from the point to its closest center. Instead of dividing by the average within-cluster distance, we could divide by the minimum or maximum within-cluster distance. If we use the latter approach, then then it would take $O(nk)$ operations to find the quality of a clustering.

3.5 Minimal Subset Quality

The user may be interested in finding the best or worst clusterable parts of a given clustering. The user may only be satisfied when all clusters are well separated from one another, or she may like to know which parts of the clustering provide the most reliable information. To

enable this type of study, given any quality measure m , we look at m over all subsets of clusters.

To find the most clusterable part of a clustering C , we can compute $\max_{S \subseteq C, |S| \geq 2} m(S, d)$. Similarly, to find the least clusterable part, we can find the subset of clusters of C that gives the lowest value of $m(S, d)$.

This approach can also be used to get variations of clustering quality measures. Some quality measures, like separability, are effected by the minimal distance between two clusters. Other measures, like standard variance ratio, can be made arbitrarily good by moving a single cluster far from all other clusters. Given any quality measure, we can modify the measure to make it sensitive to the worst section of the clustering. That is, given any quality measure m , we can define a clustering measure $m'(C, d) = \min_{S \subseteq C, |S| \geq 2} m(S, d)$. This variation is suitable for standard variance ratio and additive margin.

3.6 Clusterability and Clustering Quality

In Chapter 2, we saw notions of clusterability where the clusterability of a data set is measured by the quality of some optimal clustering. Here we present an alternative perspective at notions of clusterability.

We can say that a data set is well-clusterable if there exists a clustering that has sufficiently good clustering quality. That is, given clustering quality measure m , we can say that data set X with distance function d is well-clusterable if there exists a clustering C such that $m(C, d) \geq \alpha$. The question then becomes on how to set α . Some measures of clusterability lend themselves to some values of α . For example, when standard variance ratio is greater than 1, then the between-cluster variance is greater than the within-cluster variance. When relative margin is $\frac{1}{y}$, then on average, each point is y times closer to its center than to the next closest center. Experimental results pertaining to specific applications can also be used

to find values of α . In particular, by experimenting with random clusterings, we can find ranges that indicate poor clustering quality. Given quality measures with which it takes polynomial time to evaluate the quality of a clustering (such as the notions presented in this chapter), it is promising that we can define useful and computationally easier notions of clusterability.

3.7 Conclusions

We proposed four axioms of clustering quality measures, and demonstrated that all four axioms are non-redundant. We also introduced clustering quality measures that satisfy these axioms. These measures apply in different situations and have different properties. Separability and generalized variance ratio work with loss-based clustering. Relative and additive margin are used to evaluate the quality of center-based clusterings. For any loss function \mathcal{L} , generalized variance ratio with respect to \mathcal{L} conforms with \mathcal{L} . On the other hand, separability with k -means does not conform with k -means. Separability with k -means is sensitive to the minimal separation between clusters, while standard variance ratio and additive margin are not. We presented techniques for modifying quality measures to change their sensitivity to the best or worst clusterable parts of a clustering.

Much interesting research remains to be done. It would be interesting to see what kind of properties follow from the axioms of clustering quality measures. The ideas for alternative definitions of clusterability proposed in Section 3.6 could be developed further. In the practical domain, it would be interesting to see how the clustering quality measures presented here, and their variations, perform in practice. We hope that an interplay of theory and practice will develop as practitioners examine which measures are most useful under different circumstances, and theoreticians refine and add to the theory of clustering quality measures.

Chapter 4

Conclusions

We discussed two related but different concepts in clustering: clusterability and clustering quality. We presented three notions of clusterability from the literature and analyzed their computational complexity. We found that these notions of clusterability are inconsistent with each other. That is, for each pair of notions, there are data sets that are well-clusterable by one of the notions and poorly-clusterable by the other notion, to arbitrary degrees. Our analysis shows that worst pair ratio is the strongest of these notions of clusterability, that is, good clusterability by this notion implies good clusterability by the other notions. We also introduced a new notion of clusterability, based on point perturbation, as well as variations of this notion.

In the second part of the thesis, we proposed four axioms of clustering quality measures, and demonstrated that all four axioms are non-redundant. We also introduced clustering quality measures that satisfy these axioms. Most of these measures of clustering quality are similar to some notions of clusterability, which highlights the close relationship between these concepts. We illustrate that these measures vary on a number of dimensions, such as loss conformity and sensitivity to the minimal separation between clusters. We also suggest variations of each clustering quality measure.

This work opens many directions for future research. In the practical domain, it would be interesting to see how the clustering quality measures presented here perform in practice. Through experiments with real data sets, the ideas for alternative definitions of clusterability proposed in Section 3.6 could be developed further. In the theoretical realm, we could evaluate whether the set of axioms for clustering quality measures is complete, that is, whether all functions that do not make good clustering functions fail to satisfy at least one of the axioms. It would also be interesting to explore what other properties follow from the axioms of clustering quality measures. For clusterability, we could explore the hardness of approximating the notions of clusterability presented in Chapter 2. In addition, it would be interesting to explore axiomatization of clusterability and clustering functions using the ideas developed in this thesis.

Bibliography

- [1] T. Asano, B. Bhattacharya, M. Keil, and F. F. Yao. “Clustering algorithms based on minimum and maximum spanning trees.” *Proceedings of the 4th Annual Symposium on Computational Geometry*, 1988, pp. 252-257.
- [2] Shai Ben-David. “A framework for statistical clustering with constant time approximation algorithms for K -median and K -means clustering.” *Machine Learning*, 66:243-257, 2006.
- [3] Shai Ben-David, Nadav Eiron, Hans-Ulrich Simon. “The Computational Complexity of Densest Region Detection.” *J. Comput. Syst. Sci.* 64(1): 22-47 (2002).
- [4] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. “Clustering Large Graphs via the Singular Value Decomposition.” *Machine Learning*, 56:9-33, 2004.
- [5] S. Epter, M. Krishnamoorthy, and M. Zaki. “Clusterability detection and initial seed selection in large datasets.” Technical Report 99-6, Rensselaer Polytechnic Institute, Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY 12180, 1999.
- [6] Teofilo F. Gonzalez. “Clustering to minimize the maximum intercluster distance.” *Theoret. Comput. Sci.* 38 (1985), no. 2-3, 293–306.
- [7] Jon Kleinberg. “An Impossibility Theorem for Clustering.” *Advances in Neural Information Processing Systems (NIPS)* 15, 2002.

- [8] S. P. Lloyd. “Least squares quantization in PCM.” Special issue on quantization, *IEEE Trans. Inform. Theory*, 28:129-137, 1982.
- [9] Marina Meila. “Comparing clusterings – an information based distance.” *Journal of Multivariate Analysis archive*, 98, 5(2007):873-895.
- [10] G. Pisier, “Remarques sur un résultat non publié de B. Maurey.” *Séminaire d'Analyse Fonctionnelle*. 1980-1981, École Polytechnique, Centre de Mathématiques, Palaiseau, V.1V.12, 1981.
- [11] Narasimhan, M., Jojic, N., Bilmes, J. “Q-clustering.” *Neural Information Processing Systems (NIPS)*. Vancouver, 2005.
- [12] 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.
- [13] Bin Zhang. “Dependence of Clustering Algorithm Performance on Clustered-ness of Data.” Technical Report, 20010417. Hewlett-Packard Labs, 2001.