Clusterability: A Theoretical Study

Margareta Ackerman
University of Waterloo
Waterloo ON, Canada
mackerma@uwaterloo.ca

Shai Ben-David
University of Waterloo
Waterloo ON, Canada
shai@cs.uwaterloo.ca

Abstract

We investigate measures of the clusterability of data sets. Namely, ways to define how 'strong' or 'conclusive' is the clustering structure of a given data set. We address this issue with generality, aiming for conclusions that apply regardless of any particular clustering algorithm or any specific data generation model.

We survey several notions of clusterability that have been discussed in the literature, as well as propose a new notion of data clusterability.

Our comparison of these notions reveals that, although they all attempt to evaluate the same intuitive property, they are pairwise inconsistent.

Our analysis discovers an interesting phenomenon; Although most of the common clustering tasks are NP-hard, *finding a close-to-optimal clustering for well clusterable data sets is easy (computationally)*. We prove instances of this general claim with respect to the various clusterability notions that we discuss.

Finally, we investigate how hard it is to determine the clusterability value of a given data set. In most cases, it turns out that this is an NP-hard problem.

1 Introduction

Clustering is at the same time a very basic and an immensely useful task. However, in spite of hundreds of clustering papers being published every year, its theoretical foundations are distressingly meager. Clearly, it is very difficult to develop a theory of clustering at a level of generality that will make it relevant across different applications and algorithmic approaches. In this paper we try to take a step in that direction by investigating possible formalizations of the central and intuitive notion of *clusterability* of data sets.

The aim of clustering is to uncover meaningful partitions in data; however, not all data sets have meaningful partitions. Clusterability is a measure of clustered structure in a data set. So far, clusterability has been used only peripherally and with no theoretical support. We initiate a theoretical study of clusterability. We address this issue with generality, aiming for conclusions that apply regardless of any particular clustering algorithm or any specific data generation model.

We survey several notions of clusterability that have been discussed in the literature (some less explicitly than others), as well as propose a new notion of data clusterability. It turns out that, while all of these notions aim to capture the same intuitive concept, they are pairwise incompatible. For any pair of different notions there are data sets that are well clusterable by one notion but are poorly clusterable with respect to the other notion.

Our analysis of these notions gives rise to an interesting computational phenomenon; *Well clusterable data sets are feasibly clusterable.* We prove such claims for data sets that are well clusterable with respect to the clusterability notions we discuss. These results project on the fundamental question of evaluating the relevance to "real life" of the theory of worst-case-complexity. The ultimate strengthening of our result amounts to stating that "if a data set is hard to cluster then it is does not have a meaningful clustering structure", or that the hard cases that render clustering NP-hard are the inputs we don't care to cluster[1].

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

[1] This may resemble the situation with the basic Propo-

Finally, we investigate how hard is it to determine the clusterability of a data set. The hardness of determining clusterability has practical implications since notions of clusterability (at least the ones presented here) can be used to determine the difficulty of finding a good clustering. For each notion, we find the hardness of determining whether the clusterability of a data set exceeds a given threshold. In most cases, it turns out that this is an NP-hard problem.

We begin by presenting our new notion of clusterability, showing that data sets that are clusterable according to that notion are computationally feasible to cluster well. Next, we illustrate this phenomenon using previously proposed notions of clusterability, proving new results and and showing how previous results support our hypothesis. In Section 4, we discuss the pairwise comparison of notions of clusterability. Finally, in Section 5, we investigate how hard it is to determine the clusterability value of a given data set.

2 Framework and definitions

A *k-clustering* of data set X is a k-partition of X, that is, a set of k non-empty, disjoint subsets of X such that their union is X. A *clustering* of X is a k-clustering of X for some $k \geq 1$. For $x, y \in X$ and clustering C of X, $x \sim_C y$ whenever x and y are in the same cluster with respect to C, and $x \not\sim_C y$, otherwise.

A *notion of clusterability* is a function that takes a data set $X \subseteq \mathbf{R}^m$, and returns a real value.[2] This function is suppose to represent how 'strong' or 'conclusive' is the clustering structure of the data set.

A clustering $C = \{X_1, X_2, \ldots, X_k\}$ of $X \subseteq \mathbf{R}^m$ is *center-based* if there exist points $c_1, c_2, \ldots, c_k \in \mathbf{R}^m$, such that for all i, for all $x \in X_i$ and all $j \neq i$, $\|x - c_i\| \leq \|x - c_j\|$. The set of such points $c_1, \ldots, c_k$ is called a set of *centers* for the clustering C. The Voronoi partition induced by the centers of a clustering coincides with that clustering partition. While a partition may not have a set of centers inducing it, center-based clusterings always do.

Given a loss function $\mathcal{L}$, we let

$$OPT_{\mathcal{L},k}(X) = \min\{\mathcal{L}(C) \mid C \text{ a } k\text{-clustering of } X\},$$

the loss of a k-clustering of X that minimizes $\mathcal{L}$.

sitional Satisfiability problem, that, in spite of being the prime example of an NP-hard problem, has recently been the focus of booming industrial developments of SAT solvers that efficiently solve many large-scale practical problems.

[2]Some of our results extend beyond Euclidean space. For clarity of exposition, we have chosen to use Euclidean space throughout.

3 Clustering of well-clusterable data

Our analysis of notions of clusterability gives rise to an interesting phenomenon: *Well-clusterable data sets are computationally easy to cluster.*

Our first demonstration of this phenomenon is through a new notion of clusterability. We present a new notion of clusterability for center-based clustering, and show that using this notion, whenever a data set is well-clusterable, a provably near-optimal clustering can be computed efficiently. Next, we show that this phenomenon extends to previously known notions of clusterability, proving new results, as well as showing how previous results support our hypothesis.

The notions of clusterability that we explore fall into two broad categories. One of these categories is based on the concept of a clustering-quality measure, which is function that takes a clustering and returns a real number indicating how good or cogent is the clustering. Note that a clustering-quality measure evaluates the quality of a specific clustering, whereas a notion of clusterability evaluates the clustering tendency of a data set. For a study of clustering-quality measures, see (Ackerman and Ben-David, 2008). Given a clustering-quality measure m, we can define the clusterability of a data set X to be the optimal quality of a clustering of X. Such notions treat the clusterings that optimize clustering-quality measures as the optimal clusterings.

The second category of notions of clusterability are ones defined with respect to a clustering loss (or, objective) function used to drive clustering algorithms (such as k-means or k-median). These clusterability notions are often used in settings where the optimal clustering is defined to be the clustering optimizing such a loss function. We will discuss a few different ways of formalizing clusterability of this type.

3.1 Center perturbation clusterability

We introduce a new notion of clusterability aiming to capture the clustering robustness to center perturbations. This notion provides a distinctly different perspective at clusterability evaluation than previous notions. Center perturbation clusterability is used in conjunction with loss functions whose optimal clusterings are center-based.

Consider what happens if the centers of a center-based clustering are slightly perturbed. Are points still going to be closer to the perturbed centers of their clusters? If we re-cluster the data using the perturbed centers, how much does the loss of the clustering change? If the optimal clustering is "good", we expect such change to have little effect on clustering loss. That is, if the data

set is well-clusterable, the optimal clustering should be robust to (small) center perturbations.

ϵ-close Two center-based clusterings, C and C' of X, are *ϵ-close*, if there exist centers $c_1, c_2, \ldots, c_k$ of C, and centers $c'_1, c'_2, \ldots, c'_k$ of C', such that for all $i \leq k$, $\|c_i - c'_i\| \leq \epsilon$.

Center Perturbation Clusterability A data set X is *(ϵ, δ)-CP clusterable for k* (for $\epsilon, \delta \geq 0$), if for every clustering C of X that is ϵ-close to some optimal k-clustering of X, $\mathcal{L}(C) \leq (1+\delta)OPT_{\mathcal{L},k}(X)$.

We now prove that whenever data is well-clusterable by center perturbation, a provably near-optimal clustering can be computed efficiently, where an optimal clustering is a clustering minimizing the loss function. This is in contrast with the situation for arbitrary input, where, for more interesting loss functions (like k-means) optimal solutions are NP-hard to approximate.

Let $rad(X)$ denote the radius of the minimum hypersphere that contains all the points in X (we use the radius of X to normalize the measure to ensure scale invariance).

Theorem 1 *Given a data set $X \subseteq \boldsymbol{R}^m$ on n points, there exists an algorithm such that, for every fixed $k \geq 2$ and $\delta \geq 0$, if X is $(\frac{rad(X)}{\sqrt{\ell}}, \delta)$-CP clusterable for k, then the algorithm runs in time polynomial in n, and outputs a clustering C of X with at most k clusters, such that*

$$\mathcal{L}(C) \leq (1+\delta)OPT_{\mathcal{L},k}(X).$$

Moreover, this result holds for any loss function $\mathcal{L}$ where all optimal clusterings are center-based (an optimal k-clustering is a k-clustering minimizing $\mathcal{L}$).

We present an algorithm for finding a clustering that is ϵ-close to an optimal clustering. The algorithm is based on an algorithm by (Ben-David et al., 2002). If we know the (ϵ, δ)-CP clusterability of X, then we can lower bound the quality of the clustering found by the algorithm.

Let an *ℓ-sequence* denote a collection of ℓ elements of X (not necessarily distinct). The algorithm iterates through all k-tuples of ℓ-sequences. For each such tuple, it finds the clustering induced by the centers of mass of the ℓ-sequences. It then chooses the clustering with minimal loss.

Algorithm 1 *Finding near optimal clusterings*
INPUT: A data set X, $k \geq 1$, $\ell \geq 1$.
OUTPUT: Outputs a clustering C_A of X such that $\mathcal{L}(C_A) \leq \min\{\mathcal{L}(C) \mid C \in \mathcal{C}\}$, where $\mathcal{C}$ is the set of all k-clusterings of X that are $\frac{rad(X)}{\sqrt{\ell}}$-close to any optimal k-clustering of X.

1. *$C_A = \emptyset$*
2. *for each k-tuple of ℓ-sequences;*
   (a) *find the centers of mass of the ℓ-sequences, call this set S*
   (b) *find the clustering $\hat{C}$ that S induces on X*
   (c) *if $C_A = \emptyset$ or $\mathcal{L}(\hat{C}) < \mathcal{L}(C_A)$ then set $C_A = \hat{C}$*
3. *return C_A*

To prove Theorem 1, we use the following result by Maurey.

Theorem 2 (Maurey, 1981) *For any fixed $\ell \geq 1$ and each x' in the convex hull of X, there exist $x_1, x_2, \ldots, x_\ell \in X$ such that $\|x' - \frac{1}{\ell}\sum_{i=1}^{\ell} x_i\| \leq \frac{rad(X)}{\sqrt{\ell}}$.*

Note that $x_1, x_2, \ldots, x_\ell$ are not necessarily distinct from each other. We now prove Theorem 1.

Proof of Theorem 1

Proof By Maurey's result, there is a clustering, $\hat{C}$, examined by Algorithm 1, that is $\frac{rad(X)}{\sqrt{\ell}}$-close to an optimal clustering of X. Since Algorithm 1 selects the minimal loss clustering of the ones it reviews, $\mathcal{L}(C_A) \leq \mathcal{L}(\hat{C})$. Since $\hat{C}$ is $\frac{R}{\sqrt{\ell}}$-close to an optimal clustering of X, and X is $(\frac{R}{\sqrt{\ell}}, \delta)$-CP clusterable, $\mathcal{L}(C_A) \leq \mathcal{L}(\hat{C}) \leq (1+\delta)OPT_{\mathcal{L},k}(X)$. The running time of Algorithms 1 is $O(kmn^{\ell k+1})$. To see that, observe that there are $O(n^{\ell k})$ k-tuples of ℓ-sequences, and that for each k-tuple of ℓ-sequences the algorithm does $O(kmn)$ operations. Note that, since a clustering induced by k centers has at most k clusters, Algorithm 1 returns a clustering with no more than k clusters (for most common loss functions, including k-means, any subdivision of clusters improves the loss of a clustering).

3.2 Worst pair ratio clusterability

Worst pair ratio is an example of a notion of clusterability that is based upon a clustering-quality measure. Let the minimum distance between two points in different clusters of a clustering C be the *split* between the two clusters. Let the maximum distance between two points within a cluster in C be the *width* of the cluster. We can then evaluate the quality of a clustering by its split over width ratio. The clustering-quality

measure *worst pair ratio* is

$$WPR(C, X) = \frac{split_C(X)}{width_C(X)}.$$

Note that worst pair ratio clusterability is based on a definition by (Epter et al.).

Worst Pair Ratio Clusterability The *worst pair ratio* of X with respect to k is

$$WPR_k(X) = \max\{WPR(C, X) \mid C \text{ a } k\text{-clustering of } X\}.$$

We now prove that good worst pair ratio clusterability implies that a good clustering can be efficiently computed. In this case, a good clustering is one that is good according to the underlying quality measure, $WPR(C, X)$. In particular, we show that when worst pair ratio clusterability is sufficiently high, then the clustering that optimizes the worst pair ratio quality measure can be easily found.

First, we show that there is at most one k-clustering of a data set with split strictly greater than width.

Lemma 1 *If there exists a k-clustering C of X, for $k \geq 2$, such that $width_C(X) < split_C(X)$, then there is only one such clustering.*

Proof Assume that there are two distinct clusterings C and C' of X, each with exactly k non-empty clusters, such that $width_C(X) < split_C(X)$ and $width_{C'}(X) < split_{C'}(X)$. If $split_C(X) = split_{C'}(X)$, then $C = C'$, since each pair of points belong to the same cluster if and only if their distance is less than the split of the clustering. Assume, without loss of generality, that $split_C(X) < split_{C'}(X)$. Then every pair of points that belong to the same cluster in C also belong to the same cluster in C'. In addition, there is a pair of points that belong to the same cluster in C' but not in C (merging two clusters in C). So C' has fewer non-empty clusters than C, thus it is not a k-clustering.

Theorem 3 *Given a data set X where $WPR_k(X) \geq 1$ for some $k \geq 2$, we can find the k-clustering C with the maximum split over width ratio in $O(n^2 \log n)$ operations, where $n = |X|$.*

Proof Let C be a k-clustering that maximizes $WPR(C, X)$, over all k-clusterings of X. Then $width_C(X) < split_C(X)$. By Lemma 1, C is the unique clustering with the width strictly small than the split of the clustering.

We can run the single linkage algorithm to recover C. That is, sort the pairs of points in X based on pairwise distances, and put pairs of points in the same clusters, starting with the pair with minimal distance and going up the list until exactly k clusters are formed. Since $width_C(X) < split_C(X)$, the procedure terminates, finding C, when all edges of length at most $width_C(X)$ have been marked as within cluster edges. This takes $O(n^2 \log n)$ operations.

3.3 Separability clusterability

Separability is another notion of clusterability that evaluates the clusterability of a data set with respect to a loss function, although it does that in a distinctly different manner than center perturbation. Separability captures how sharp is the drop in the loss function when moving from a $(k-1)$-clustering to a k-clustering. This notion was introduced by (Ostrovsky et al., 2006).

Separability was defined with respect to the k-means loss function (although it clearly applies with other loss functions as well), $k\text{-means}(C) = \sum_{i=1}^{k} \sum_{x \in X_i} \|x - \textit{center-mass}(X_i)\|^2$, where $\textit{center-mass}(X) = \frac{1}{|X|} \sum_{x \in X} x$.

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

For convenience, we define $S_k(X)$ to be the smallest ϵ such that X is (k, ϵ)-separable. The range of separability is $[0, 1)$, and a data set has better separability than another data set if it is separable for smaller ϵ.

Ostrovsky et al. prove that given data with good separability clusterability, it is easy to find a clustering with good k-means loss.

Theorem 4 (Theorem 3.4, (Ostrovsky et al., 2006)) *Given a $(2, \epsilon^2)$-separable data set $X \subseteq \boldsymbol{R}^m$, we can find a 2-clustering with k-means loss at most $\frac{OPT_2(X)}{1-\rho}$ with probability at least $1 - O(\rho)$ in time $O(nm)$, where $\rho = \Theta(\epsilon^2)$ and $n = |X|$.*

For arbitrary k, (Ostrovsky et al., 2006) provide a similar result.

3.4 Variance ratio clusterability

Variance ratio evaluates the clusterability of a data set with respect to a normalized loss function. As such, it belongs to both categories of clusterability notions examines so far - notions defined in terms of loss functions, and ones that set clusterability to the optimal quality of a clustering.

Variance ratio measures the ratio of the variance between clusters over the variance of points within clusters. This notion was presented by (Zhang, 2001). Recall that the variance of X is $\sigma^2(X) =$

$\frac{1}{|X|}\sum_{x\in X}\|x - \textit{center-mass}(X)\|^2$. Consider a clustering $C = \{X_1, X_2, \ldots, X_k\}$. Let $p_i = \frac{|X_i|}{|X|}$. Let $B_C(X) = \sum_{i=1}^{k} p_i \|\textit{center-mass}(X_i) - c\|^2$ denote the *between-cluster variance* of a clustering C and $W_C(X) = \sum_{i=1}^{k} p_i \sigma^2(X_i)$ the *within-cluster variance* of a clustering C. The clustering-quality measure *variance ratio* is

$$VR(C, X) = \frac{B_C(X)}{W_C(X)}.$$

Variance Ratio Clusterability The *variance ratio* of X for k is

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

where $\mathcal{C}$ is the set of k-clusterings of X.

Observe that $\sigma^2(X) = W_C(X) + B_C(X)$ and $W_C(X)$ is the loss function that k-means minimizes divided by n. In addition, $VR_k(X) = \frac{\sigma^2(X) - W_C(X)}{W_C(X)}$. Since $\sigma^2(X)$ is constant over all clusterings of X, a clustering that optimizes the k-means loss function also optimizes variance ratio. For this reason, we can view variance ratio as a normalization of the k-means loss function. We let $W_k(X) = W_C(X)$ and $B_k(X) = B_C(X)$, where C is a clustering that optimizes k-means objective function.

The range of variance ratio is $[0, \infty)$ and higher values of variance ratio indicate better clusterability.

We prove that when variance ratio is good for two clusters, a provably near-optimal clustering can be efficiently computed. In this context, an optimal clustering is one that optimizes the variance ratio clustering-quality measure, or, equivalently, is it a clustering that optimizes the k-means loss.

We make use of the result of (Ostrovsky et al., 2006) stated in Theorem 4.

First, we show that variance ratio and separability are equivalent for $k = 2$.

Lemma 2 *$VR_2(X) = \frac{1}{S_2(X)} - 1$ for any data set X.*

Proof Let $n = |X|$. We know that $\sigma^2(X) = W_2(X) + B_2(X)$, and $W_2(X) = \frac{OPT_2(X)}{n} = S_2(X)\sigma^2(X)$. Thus, $VR_2(X) = \frac{B_2(X)}{W_2(X)} = \frac{\sigma^2(X) - W_2(X)}{W_2(X)} = \frac{\sigma^2(X) - S_2(X)\sigma^2(X)}{S_2(X)\sigma^2(X)} = \frac{1}{S_2(X)} - 1$.

Combining Lemma 2 with Theorem 4, we get that a data set with good variance ratio clusterability is easy to cluster well.

Corollary 1 *Given a data $X \subseteq \boldsymbol{R}^m$, we can find a 2-clustering with k-means loss at most $\frac{OPT_2(X)}{1-\rho}$ with probability at least $1 - O(\rho)$ in time $O(nm)$, where $\rho = \Theta\left(\frac{1}{(VR_2(X)+1)^2}\right)$.*

3.5 Clusterability Assuming a Target Clustering

The work of (Balcan et al., 2008), and (Balcan et al., 2009) gives further evidence that well-clusterable data is easy to cluster. In both papers, the authors assume the existence of an unknown *target clustering* that the user hopes to uncover or approximate.

3.5.1 Strict separation clusterability

Balcan, Blum, and Vempala look for properties of data sets that enable efficient clustering, *up to a list or tree*. That is, the goal is to find a list of clusterings containing all the clusterings that satisfy a certain property, or, alternately, a tree whose prunings include all such clusterings. In contrast, in our work, we have found properties of data sets such that a *single* good clustering can be efficiently computed. In particular, good clusterability according to center-perturbation, separability, or variance ratio, means that a clustering that approximates the optimal loss for certain loss functions can be efficiently computed. For data sets with sufficiently good worst pair ratio, a single clustering optimizing the underlying clustering-quality measure is easy to find.

Balcan et al. present the following property of a clustering.

Strict Separation Property A clustering $C = \{C_1, C_2, \ldots, C_k\}$ over data set X satisfies *strict separation* if every point is closer to every point in its cluster, than to any point in another cluster. That is, for any i and any $x \in C_i$, given any point $y \in C_i$ and $z \notin C_i$, x strictly closer to y than to z.

We can convert this property to a notion of clusterability, as follows.

Strict Separation of a Clustering The *Strict Separation* of a clustering C over X is

$$StrS(C, X) = \max_{x\in X} \frac{\max_{x\sim_C y}\|x - y\|}{\min_{x\not\sim_C y}\|x - y\|}.$$

Strict Separation Clusterability The *Strict Separation* of X is

$$StrS(X) = \min\{StrS(C, X) \mid C \text{ a clustering of } X\}.$$

(Balcan et al., 2008) prove the following result, supporting the hypothesis that well-clusterable data is easy to cluster.

Theorem 5 (Theorem 2, (Balcan et al., 2008)) *If $StrS(X) \geq 1$, then we can efficiently construct a tree such that every clustering C where $StrS(C, X) \geq 1$ is a pruning of this tree.*

3.5.2 Target clusterability

In "Approximate clustering without approximation," (Balcan et al., 2009) prove that clustering is easier if we permit the assumption that approximate solutions of certain clustering loss functions approximate the unknown target clustering. This assumption can be used to formulate a notion of clusterability. We now present their results (rephrasing them in the context of clusterability).

Given the target clustering C, let the *error* of clustering C' be the number of points that it labels differently than the target clustering (under the best permutation of labels.)

$(\mathcal{L}, c, \epsilon)$-Target Clusterability A data set X satisfies $(\mathcal{L}, c, \epsilon)$-*target* clusterability if any near optimal k-clustering C, where $\mathcal{L}(C) \leq OPT_{\mathcal{L},k}(X)$, has error at most ϵ (with respect to the target clustering).

Target clusterability resembles the notions we examine that make use of loss functions. However, distinct from these notions, target clusterability views the target clustering as the optimal clustering, which does not necessarily optimize the loss function (although it does approximate it).

(Balcan et al., 2009) prove that, for the loss functions k-means, k-median, and min-sum, well-clusterability by $(\mathcal{L}, c, \epsilon)$-target clusterability implies that the underlying data set is computationally easy to cluster. We present their result for the k-means loss function.

Theorem 6 (Theorem 10, (Balcan et al., 2009)) *If a data set satisfies $(k\text{-means}, 1+\alpha, \epsilon)$-target clusterability, then we can efficiently find a clustering that is $O(\epsilon/\alpha)$-close to the target clustering.*

4 Comparison of notions of clusterability

We have performed a pairwise comparison of notions of clusterability, finding that no two are equivalent. In particular, we analyzed separability, variance ratio, worst pair ratio, and center perturbation. For illustration purposes, we present the detailed comparison between variance ratio and separability. We omit the remaining proofs due to lack of space, noting that the proofs showing that no pair of notions are equivalent are not difficult to obtain.

The comparison of variance ratio and separability reveals that good clusterability by separability implies that variance ratio clusterability is good, however, good clusterability according to variance ratio does not imply good clusterability by separability.

Lemma 3 $W_k(X) = S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X)$, *for any $k \geq 2$ and $X \subseteq \mathbf{R}^m$,*

Proof The result holds for $k = 2$, since $W_2(X) = \frac{OPT_{k\text{-means},2}(X)}{|X|} = \frac{S_2(X)|X|\sigma^2(X)}{|X|} = S_2(X)\sigma^2(X)$. Assume that the result holds for all $j < k$. Then,

$$\begin{aligned} W_k(X) &= \frac{1}{|X|} OPT_{k\text{-means},k}(X) \\ &= \frac{1}{|X|} S_k(X) OPT_{k\text{-means},k-1}(X) \\ &= \frac{1}{|X|} S_k(X)|X|W_{k-1}(X) \\ &= S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X) \end{aligned}$$

Theorem 7 $VR_k(X) = \frac{VR_{k-1}(X)+1}{S_k(X)} - 1$, *for $k \geq 3$ and $X \subseteq \mathbf{R}^m$.*

Proof By Lemma 3, $W_k(X) = S_2(X)S_3(X)\cdots S_k(X)\sigma^2(X)$. Then,

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

By the above result, we can show that good clusterability by separability implies that variance ratio clusterability is good.

Theorem 8 $VR_k(X) \geq \frac{1}{S_k(X)} - 1$ *for any $k \geq 2$ and data set $X \subseteq \boldsymbol{R}^m$.*

Proof By Theorem 7, for any $k \geq 3$,

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

Since $VR_{k-1}(X) \geq 0$, this implies that $VR_k(X) \geq \frac{1}{S_k(X)} - 1$. For $k = 2$, the result follows from Lemma 2.

We now show that good clusterability according to variance ratio does not imply good clusterability by separability.

Theorem 9 *For any $\epsilon < 1$, there is a data set X and a $k \geq 3$ such that $S_k(X) \geq \epsilon$ and $VR_k(X)$ is arbitrarily high.*

Proof It can be shown that there exists a k such that $S_{k-1}(X) \geq \epsilon$ for any $\epsilon < 1$. Choose one such data set. Then add another point sufficiently far away from all other points in the data set, to increase the between-cluster variance, making $VR_k(X)$ arbitrarily high. Place the 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 in an optimal k-means 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, and therefore $S_k(X) = S_{k-1}(X) \geq \epsilon$.

5 Computational complexity of clusterability

We analyze the computational complexity of determining the clusterability of a data set. To the best of our knowledge, this is the first computational complexity analysis of notions of clusterability. The hardness of determining clusterability has practical implications since, as shown in Section 3, notions of clusterability (at least the ones presented here) can be used to determine the difficulty of finding a good clustering.

Our results show for when worst pair ratio is sufficiently good, worst pair ratio clusterability can be found in polynomial time. However, it is easy to see that worst pair ratio is very sensitive to noise and outliers, and thus often assigns low clusterability to intuitively well-clusterable data sets. Therefore, good clusterability for worst pair ratio implies particularly clear clustering structure. For separability and variance ratio clusterability, which can detect clustered structure in a wider range of circumstances, the problem of determining the degree of clusterability is NP-hard.

5.1 Complexity of separability

We prove that determining whether the separability clusterability of a given data set exceeds a given threshold is an NP-hard problem.

Theorem 10 *Given $X \subseteq \mathbf{R}^m$, integer $k \geq 2$, and $0 < \epsilon < 1$, it is NP-hard to determine whether X is (k, ϵ)-separable.*

Proof The decision version of the k-means problem is as follows: Does there exist a k-clustering of X with k-means loss at most v? This problem is NP-complete for $k \geq 2$ (Drineas et al., 2004).

A data set X is (2,ϵ)-separable if $\frac{OPT_{k\text{-means},2}(X)}{OPT_{k\text{-means},1}(X)} \leq \epsilon$. Suppose that we could determine whether any set $X \subseteq \mathbf{R}^m$ is (2,ϵ)-separable for any $0 < \epsilon < 1$ in polynomial-time. Then since $OPT_{k\text{-means},2}(X) \leq \epsilon OPT_{k\text{-means},1}(X)$, and $OPT_{k\text{-means},1}(X) = |X|\sigma^2(X)$ can be found in polynomial-time, we can determine whether $OPT_{k\text{-means},2}(X) \leq \mu$ for any μ by checking if X is (2,ϵ)-separable for $\epsilon = \frac{\mu}{OPT_{k\text{-means},1}(X)}$. But since determining if $OPT_{k\text{-means},2}(X) \leq \mu$ for any arbitrary $\mu > 0$ is NP-hard, determining if X is (2,ϵ)-separable is NP-hard.

To show that the problem is NP-hard for any $k \geq 3$, we reduce the problem for $k = 2$ to the problem for $k \geq 3$. 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 is its own cluster in the k-means optimal clustering. Then in any k-means optimal clustering, the remaining 2 clusters are an optimal 2-means solution for the original data set.

5.2 Complexity of variance ratio

Determining the level of clusterability is also an NP-hard problem for the variance ratio notion of clusterability.

Theorem 11 *Given $X \subseteq \mathbf{R}^m$, $k \geq 2$, and $r > 0$, it is NP-hard to determine whether $VR_k(X) \geq r$.*

Proof We know that $\sigma^2(X) = W_k(X) + B_k(X)$. Then $VR_k(X) = \frac{B_k(X)}{W_k(X)} = \frac{\sigma^2(X) - W_k(X)}{W_k(X)} = \frac{\sigma^2(X)}{W_k(X)} - 1 = \frac{|X|\sigma^2(X)}{OPT_{k\text{-means},k}(X)} - 1$.

Thus, if we can tell whether $VR_k(X) = \frac{|X|\sigma^2(X)}{OPT_{k\text{-means},k}(X)} - 1 \geq r$ for any $r > 0$, then we can tell whether $OPT_{k\text{-means},k}(X) \leq \frac{|X|\sigma^2(X)}{r+1}$. We can find $|X|\sigma^2(X)$ in polynomial time. Also, by definition of $OPT_{k\text{-means},k}(X)$, $OPT_k(X) \leq |X|\sigma^2(X)$. Thus, by setting $r = \frac{|X|\sigma^2(X)}{v} - 1$, we can find whether $OPT_{k\text{-means},k}(X) \leq v$ for any $v > 0$. However, this problem is NP-hard for $k \geq 2$ (Drineas et al., 2004).

5.3 Complexity of worst pair ratio

We show that whenever worst pair ratio clusterability is sufficiently good, then it can be determined in

polynomial time.

Theorem 12 *Given an integer $k \geq 2$ and a data set $X \subseteq \mathbf{R}^m$ where $WPR_k(X) > 1$, we can determine the worst pair ratio of X in polynomial-time.*

Proof By Theorem 3, given a data set X on n points where $WPR_k(X) > 1$, a k-clustering C that maximizes the split over width ratio over all k-means optimal clusterings of X can be found in $O(n^2 \log n)$ operations. We can then find the split and width of C using $O(n^2)$ additional operations, thus finding $WPR_k(X)$.

6 Conclusions

In this work, we present a theoretical study of clusterability. We survey some previous notions of clusterability, and present a new notion that captures clustering robustness to center perturbations. Our comparison of these notions shows that, although they attempt to measure the same intuitive property, they are pairwise inconsistent.

Our analysis reveals an interesting property common to these notions of clusterability: when a data set is well-clusterable, it is computationally easy to find a near optimal clustering of that data set. This phenomenon occurs across provably distinct notions of clusterability, and even different notions of clustering optimality. It is intriguing to figure out how broad this phenomenon is. In particular, it is an interesting challenge to the research community to obtain similar results with other natural notions of clusterability. Such results may help understand when certain clustering algorithms perform well, and show that when they fail to produce satisfactory clusterings, it is due to insufficient clustering structure in the data.

We explore the hardness of determining the degree of clusterability of a given data set using previous notions of clusterability. Our analysis shows that, for the notions that recognize a wide range of well-clustered data, this is an NP-hard problem.

For future work, it would be interesting to explore the problem of determining clusterability in the more flexible framework of property testing. The goal of property testing is to determine whether a given object has a desired property or if the object is close to some other object which has the property. The tester is allowed to make mistakes on both positive and negative assertions, with a certain probability. For more details on property testing, see (Goldreich et al., 1998). Using this setting, we pose the following question: given a notion of clusterability, we ask how hard is it to determine whether the clusterability of a data set surpasses a given threshold or if the data set is similar to some data set that does.

References

M. Ackerman and S. Ben-David. Measures of clustering quality: A working set of axioms for clustering. In *NIPS*, 2008.

M.F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th annual ACM symposium on Theory of Computing*, pages 671–680. ACM New York, NY, USA, 2008.

M.F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the approximation. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2009.

S. Ben-David, N. Eiron, and H.U. Simon. The computational complexity of densest region detection. *Journal of Computer and System Sciences*, 64(1): 22–47, 2002.

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33, 2004.

S. Epter, M. Krishnamoorthy, and M. Zaki. Clusterability detection and initial seed selection in large datasets. The International Conference on Knowledge Discovery in Databases.

O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 165–176. IEEE Computer Society Washington, DC, USA, 2006.

B. Zhang. Dependence of Clustering Algorithm Performance on Clustered-ness of Data. 2001.