Clustering is Easy When . . . What?

Shai Ben-David
Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada
shai@uwaterloo.ca

Abstract

It is well known that most of the common clustering objectives are NP-hard to optimize. In practice, however, clustering is being routinely carried out. One approach for providing theoretical understanding of this seeming discrepancy is to come up with notions of clusterability that distinguish realistically interesting input data from worst-case data sets. The hope is that there will be clustering algorithms that are provably efficient on such "clusterable" instances. This paper addresses the thesis that the computational hardness of clustering tasks goes away for inputs that one really cares about. In other words, that "Clustering is difficult only when it does not matter"[1] (the *CDNM thesis* for short).

I would like to use the format of a workshop presentation to deviate from the conference paper style of focusing on cutting edge results. Instead, I wish to present a a critical bird's eye overview of the results published on this issue so far and to call attention to the gap between available and desirable results on this issue. I start by discussing which requirements should be met in order to provide formal support to the the CDNM thesis. I then examine existing results in view of these requirements and list the most significant unsolved research challenges in that direction.

1 Introduction

Computational complexity theory aims to provide tools for the quantification and analysis of the computational resources needed for algorithms to perform computational tasks. Worst-case complexity is by far the best known, most researched and best understood approach to computational complexity theory. In particular, NP-hardness is a worst-case-instance notion. By saying that a task is NP-hard (and assuming $P \neq NP$), we imply that for every algorithm, there exist infinitely many instances on which it will have to work hard. However, for many problems this measure is unrealistically pessimistic compared to the experience of solving them for practical instances. A problem may be NP–hard and still have algorithms that solve it efficiently for any instance that is likely to occur in practice or any instance for which one cares to find an optimal solution for.

Here, we focus on clustering tasks that are defined as discrete optimization problems. Most of those optimization problems are NP-hard. We wish to examine whether this hardness remains an issue when we restrict our attention to "clusterable data" - data for which a meaningful clustering exists (one can argue that when there is no cluster structure in a given data set, there is no point in applying a clustering algorithm to it). In other words, we wish to evaluate to what extent current theoretical work supports the "Clustering is difficult only when it does not matter" (CDNM) thesis. In this note, we provide a relatively high level view of some of the major relevant results. A more detailed version of our results can be found in [2]. For the sake of concreteness, we will focus on two popular clustering objectives, k-means and k-median.

[1] This phrase is in fact a title of a recent paper – [1].

We start this note by listing, in Section 2, what we think are requirements from notions of clusterability aiming to substantiate the CDNM thesis. In Section 3, we list various notions of clusterability that have been proposed in the context of this line of research.

We then examine the results pertaining to the proposed notions of clusterability listed above, from the perspective of those requirements. Due to the conciseness of this note, we list here only some representative of our results and refer to the full version [2] for a more complete list.

Our conclusion is that the currently available theory is still far from substantiating the CDNM thesis. In particular, while additive perturbation robustness, with any non-zero robustness parameter, gives rise to algorithms that find the optimal clusterings in time polynomial in the input size and its dimension, as far as currently published results go, non of the requirements listed above allows finding optimal clustering solutions in time polynomial in the number of target clusters, k, unless the corresponding parameters are set to values that hold only for extremely well clusterable data sets

In Section 5 we summarize these discouraging results and highlight some implied open problems and propose directions in which this line of research should, in our opinion, proceed.

2 Requirements from notions of clusterability

We begin by stating requirements that (we believe) a notion of clusterability should satisfy to be applied for supporting the "Clustering is Difficult only when it does Not Matter" (CDNM, in short) thesis. Those requirements are stated as qualitative, high level, statements.

1. *It should be reasonable to assume that most (or at least a significant proportion of) the inputs one may care to cluster in practice satisfy the clusterability notion.*

Of course, we do not have any way to guarantee that unseen practical instances will satisfy any non-trivial requirement. However, this type of consideration can serve as a way to filter out clusterability conditions that are too restrictive. Furthermore, when a good data generative model is available, one can formalize requirements pertaining to a high probably of having the generated instances satisfy the given clusterability notion.

2. *In order to support the CDNM thesis, a notion of clusterability should be such that there exist efficient algorithms that are guaranteed to find a good clustering (minimizing the objective function, or getting very close to it) for any input that satisfies that clusterability requirement.*

The next two requirements may be more debatable. They are motivated by considering practical aspects of clustering applications. Assume we do have some clusterability condition and a guarantee that the algorithm we are about to run is efficient on instances satisfying it. When we get some real input, there is no guarantee that it satisfies that clusterability condition. Since for most of the NP-hard clustering problems, there is no efficient way of measuring how far from optimal a given clustering solution is, one may not being able to protect against bad solutions.

3. *There exists an efficient algorithm for testing clusterability. Namely, given an instance (X, d), the algorithm determines whether it satisfies the clusterability requirement or not.*

A forth, somewhat orthogonal, desiderata relates to existing common clustering algorithms. Namely,

4. *Some commonly used clustering algorithm can be guaranteed to perform well (i.e., run in poly-time and find close-to-optimal solutions) on all instances satisfying the clusterability assumption.*

Requirement 4 is important if our goal is to *understand* what is happening nowadays in clustering work by providing a theoretical explanation for the success of common clustering algorithms on real data. However, even when failing it, requirement 2 may lead to the development of new clustering algorithms, which may have independent merits.

The main Open Question: *Find a notion of clusterability that satisfies the requirements above (or even just the first two).*

3 Notions of clusterability

In the past few years there have been several interesting publications along the lines described above, showing that for various notions of clusterability there are indeed algorithms that find optimal clusterings in polytime for all appropriately clusterable instances. Below is a (possibly not exhaustive) list of major notions of clusterability that have been discussed in that context[2]. Most of these definitions can be applied to any center-based clustering objective. Due to space constrains, this version omits some of the technical details of the following definitions.

1. **Perturbation Robustness:** An input data set is perturbation robust if small perturbations of it do not result in a change of the optimal clustering for that set.
   (a) Additive perturbation robustness (APR) [3][3]: An input set (X, d) is ϵ-APR if some optimal k-clustering C remains optimal for any small (additive) perturbation of this input [4].
   (b) Multiplicative perturbation robustness (MPR) [5]: An input set (X, d) is α-MPR if some optimal k-clustering C remains optimal for any small (multiplicative) perturbation of this input.
2. **Significant loss of the objective when reducing the number of clusters:**
   (a) ϵ **-Separatedness:** [6] discuss clustering w.r.t. the k-means objective. They define An input data set (X, d) is *ϵ-separated for k* if the k-means cost of the optimal k-clustering of (X, d) is less then ϵ^2 times the cost of the optimal $(k-1)$-clustering of (X, d).
   (b) **Weak Deletion Stability:** [7] An instance for k-clustering satisfies the $(1 + \alpha)$ *Weak Deletion Stability* condition if, for its optimal clustering, removing any center c_i and assigning all the points in its cluster to a different center c_j, results in an increase of cost of the clustering by a factor $\geq (1 + \alpha)$.
3. **Center stability**: [8] An instance (X, d) is *α-center stable* (with respect to some center based clustering objective Ø) if for any optimal clustering of it, all points are closer by a factor α to their own cluster center than to any other cluster center.
4. **Uniqueness of optimum**: [9] A data set is *(c, ϵ)-approximation-stable* with respect to some *target clustering C_T* if every clustering C of X whose objective cost over (X, d) is within a factor c of the objective cost of C_T (on (X, d)) is ϵ-close to C_T (w.r.t. some natural notion of between-clustering distance). This condition rules out the possibility of having two significantly different close-to-optimal-cost solutions.

4 To what extent do the notions meet the requirements listed above?

As varied as the above list of proposed notions may sound, it turns out that almost all (except for the additive perturbation robustness, which is also the only one that does not yield efficiency for large k) imply that data satisfying them is structured such that the vast majority of the data points can be assigned to compact clusters that are very widely separated (or that all but a small fraction of the clusters are such). We provide quantitative versions of this claim in Section 4.1. In fact, this common characteristic of the notions is the main feature that is being used in showing that, under such conditions, clustering can be carried out efficiently. While all of the above notions sound intuitively plausible (concrete arguments supporting that plausibility can be found in the papers presenting them), the quantitative values of the clusterability assumptions are essential for evaluating that plausibility. We show (see [2]) that the currently known results concerning these notions yield the desired efficiency of computation only when the clusterability parameters are set to values that are beyond what one might expect practical inputs to satisfy.

[2] be ware that different papers use different terminology for similar notions (and similar terminology for different notions), so my choice of terminology below is not always consistent with other publications.

[3] The definition of robustness, as well as the implied efficiency of clustering result, in [3] are particular cases of a more general definition and more general results of [4]

[4] Since this additive condition is not scale invariant, we implicitly add the assumption that the diameter of the input set, $\max_{x,y \in X} d(x, y)$, is at most 1 (otherwise the stability parameter should be multiplied by that diameter).

4.1 Computational efficiency vs realistic soundness of clusterable inputs

For all the above mentioned clusterability condition, it has been shown that when the clusterability parameters are set to sufficiently restrictive values, data satisfying those requirements allows poly-time discovery of its optimal clustering solutions. However, examining the implications of those parameter settings to the input data satisfying them, we conclude that efficiency is obtained only for rather unrealistic data setup. Typical examples of our results are (for a full list of those results see [2]):

- The values of ϵ for which ϵ -Separatedness is shown (in [6]) to allow $poly(k)$ clustering algorithms imply that, in the optimal clustering, the average distance of a point from its cluster center should be smaller than the minimal distance between distinct cluster centers by a factor of at least 200.
- The values of parameters for which (c, ϵ) approximation stability is shown (in [10]) to allow $poly(k)$ clustering algorithms imply that, in the optimal clustering, for all but an ϵ-fraction of the input points, the distance of a point to its own cluster center is smaller than its distance to the next closest center by at least 20 times the average point-to-its-cluster-center-distance.
- The values of α for which $(1+\alpha)$ weak deletion stability is shown (in [7]) to allow $poly(k)$ clustering algorithms imply that, in the optimal clustering, the vast majority of the clusters are so distant from the rest of the data points that any point outside such a cluster is further from the center of that cluster by at least $\log(k)$ times the "average radius" of its own cluster.

4.2 Efficient testability of the clusterability conditions

When it comes to testing whether a given clustering instance satisfies any of the above clusterability conditions, a key point to note is that they are all phrased in terms of condition pertaining to the optimal clustering of the given data. Finding such optimal clusterings is NP-hard. Furthermore, there exist no efficient algorithm for testing, given a data set (X, d) and a k clustering of it, C, whether C is an optimal clustering for (X, d).

4.3 Implications for common practical clustering algorithms

Among all the works surveyed in this note, only one, the results of [6], address (a feasible variant of) a practical algorithm - the popular Lloyd clustering algorithm. It would be very interesting to come up with results showing that some popular clustering algorithm (or an application of a practical approximation algorithm) efficiently yield guaranteed good quality clusterings, under some other, or more relaxed, niceness of data conditions.

5 Conclusions

For each notion of "easy clustering inputs" proposed so far, the parameter values that suffice for the currently available efficient clustering results turns out to be too strong requirements from the practical significance perspective. The current failure to support the CDNM thesis may stem from various sources. First, of course, maybe the thesis is just false. My personal belief is that, while it may very well be the case that some practical clustering tasks are indeed computationally hard for some real data instances, there are many more cases where data of practical interest does yield not-too-hard-to-find meaningful clusterings (though, of course, most of the time we have no way of knowing whether those are optimal clusterings in any formal sense of optimality).

Another explanation to the shortcomings of current results is that they may just be an artifact of the algorithms and proof techniques that we currently have. I doubt if that is indeed the case. In fact, [12] shows an NP-hardness result for center stability that almost matches the parameter values known to imply feasibility under such a condition.

I believe that part of the answer is that we have not yet discovered the appropriate notions of clusterability. The results surveyed in this paper indicate that notions of clusterability that aim to substantiate the CDNM statement should not be just a way of formalizing large between-clusters separation.

References

[1] Amit Daniely, Nati Linial, and Michael Saks. Clustering is difficult only when it does not matter. *CoRR*, abs/1205.4891, 2012.

[2] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *arXiv preprint arXiv:1501.00437*, 2015.

[3] Margareta Ackerman and Shai Ben-David. Clusterability: A theoretical study. In *AISTATS*, pages 1–8, 2009.

[4] Shai Ben-David. Alternative measures of computational complexity with applications to agnostic learning. In *TAMC*, pages 231–235, 2006.

[5] Yonatan Bilu and Nathan Linial. Are stable instances easy? In *ICS*, pages 332–341, 2010.

[6] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28, 2012.

[7] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *FOCS*, pages 309–318, 2010.

[8] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, 2012.

[9] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Clustering under approximation stability. *J. ACM*, 60(2):8, 2013.

[10] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *SODA*, pages 1068–1077, 2009.

[11] Pranjal Awasthi, Afonso Bandera, Moses Charikar, Ravishankar Krishnaswami, Soledad Voilar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. *CoRR*, Stat.ML, 2014.

[12] Shalev Ben-David and Lev Reyzin. Data stability in clustering: A closer look. *Theoretical Computer Science*, page To appear, 2014.

[13] Lev Reyzin. Data stability in clustering: A closer look. In *ALT*, pages 184–198, 2012.

[14] Shai Ben-David, Dávid Pál, and Hans-Ulrich Simon. Stability of k -means clustering. In *COLT*, pages 20–34, 2007.

[15] Ohad Shamir and Naftali Tishby. Stability and model selection in k-means clustering. *Machine Learning*, 80(2-3):213–243, 2010.