Chromatic Clustering in High Dimensional Space

Hu Ding   Jinhui Xu

Department of Computer Science and Engineering
State University of New York at Buffalo
{huding, jinhui}@buffalo.edu

1 Overview

Clustering is one of the most fundamental problems in computer science and finds applications in many different areas [2, 3, 5, 6, 11, 14, 15, 17, 19]. Most existing clustering techniques assume that the to-be-clustered data items are independent from each other. Thus each data item can "freely" determine its membership within the resulting clusters, without paying attention to the clustering of other data items. In recent years, there are also considerable attentions on clustering dependent data and a number of clustering techniques, such as correlation clustering, point-set clustering, ensemble clustering, and correlation connected clustering, have been developed [3, 6, 11, 14].

In this paper, we consider the following new type of clustering problems, called *Chromatic Clustering*, for dependent data. Let $\mathcal{G} = \{G_1, \cdots, G_n\}$ be a set of n point-sets with each $G_i = \{p^i_1, \ldots, p^i_{k_i}\}$ consisting of $k_i \leq k$ points in $\mathbb{R}^d$ space. A chromatic partition of $\mathcal{G}$ is a partition of the $\sum_{1\leq i\leq n} k_i$ points into k sets, $U_1, \cdots, U_k$, such that each U_i contains no more than one point from each G_j for $j = 1, 2, \cdots, n$, where the dimensionality d could be very high. The chromatic k-means clustering (or k-CMeans) of $\mathcal{G}$ is to find k points $\{m_1, \cdots, m_k\}$ in $\mathbb{R}^d$ space and a chromatic partition $U_1, \cdots, U_k$ of $\mathcal{G}$ such that $\frac{1}{n}\sum_j \sum_{q\in U_j} ||q - m_j||^2$ is minimized. The problem is called full k-CMeans if $k_1 = k_2 = \cdots = k_n = k$. Similarly we can define chromatic k-Median clustering (or k-CMedians) for $\mathcal{G}$. Chromatic clustering captures the mutual exclusiveness relationship among data items and is a rather useful model for various applications. Due to the additional chromatic constraint, chromatic clustering is thus expected to simultaneously solve the "coloring" and clustering problems, which significantly complicates the problem. We are able to show that the chromatic clustering problem is challenging to solve even for the case that each color is shared only by two data items.

Related works: As its generalization, chromatic clustering is naturally related to the traditional clustering problem. Due to the additional chromatic constraint, chromatic clustering could behave quite differently from its counterpart. For example, the k-means algorithms in [5, 8, 12, 13, 18] relies on the fact that all input points in a Voronoi cell of the optimal k mean points belong to the same cluster. However, such a key locality property no longer holds for the k-CMeans problem.

Chromatic clustering falls in the umbrella of clustering with constraint. For such type of clustering, several solutions exist for some variants [4, 7, 10]. Unfortunately, due to their heuristic nature, none of them can yield quality guaranteed solutions for the chromatic clustering problem. The first quality guaranteed solution for chromatic clustering was obtained recently by Ding and Xu. In [14], they considered a special chromatic clustering problem, where every point-set has exactly k points in the first quadrant, and the objective is to cluster points by cones apexed at the origin, and presented the first PTAS for constant k. The k-CMeans and k-CMedians problems considered in this paper are the general cases of the chromatic clustering problem. Very recently, Arkin *et al.* [1] considered a chromatic 2D 2-center clustering problem and presented both approximation and exact solutions.

1.1 Main Results and Techniques

In this paper, we present three main results, a constant approximation and a $(1+\epsilon)$-approximation for k-CMeans and their extensions to k-CMedians.

- **Constant approximation:** We show that given any λ-approximation for k-means clustering, it could yield a $(18\lambda + 16)$-approximation for k-CMeans. This not only provides a way for us to generate an initial constant approximation solution for k-CMeans through some k-means algorithm, but more importantly reveals the intrinsic connection between the two clustering problems.
- $(1 + \epsilon)$**-approximation:** We show that a near linear time $(1 + \epsilon)$-approximation solution for k-CMeans can be obtained using an interesting sphere peeling algorithm. Due to the lack of locality property in k-CMeans, our sphere peeling algorithm is quite different from the ones used in [5,18], which in general do not guarantee a $(1+\epsilon)$-approximation solution for k-CMeans as shown by our first result. Our sphere peeling algorithm is based on another standalone result, called *Simplex Lemma*. The simplex lemma enables us to obtain an approximate mean point of a set of unknown points through a grid inside a simplex determined by some partial knowledge of the unknown point set. A unique feature of the simplex lemma is that the complexity of the grid is *independent of the dimensionality*, and thus can be used to solve problems in high dimensional space. With the simplex lemma, our sphere peeling algorithm iteratively generates the mean points of k-CMeans with each iteration building a simplex for the mean point.
- **Extensions to k-CMedians:** We further extend the idea for k-CMeans to k-CMedians. Particularly, we show that any λ-approximation for k-medians can be used to yield a $(3\lambda+2)$-approximation for k-CMedians. With this and a similar sphere peeling technique, we obtain a $(1 + \epsilon)$-approximation for k-CMedians.

References

1. Esther M. Arkin, Jos Miguel Daz-Bez, Ferran Hurtado, Piyush Kumar, Joseph S. B. Mitchell, Beln Palop, Pablo Prez-Lantero, Maria Saumell, Rodrigo I. Silveira: Bichromatic 2-Center of Pairs of Points. LATIN 2012: 25-36
2. David Arthur, Sergei Vassilvitskii: ”k-means++: the advantages of careful seeding”. *SODA 2007: 1027-1035*
3. Nikhil Bansal, Avrim Blum, Shuchi Chawla: ”Correlation Clustering”.*Machine Learning 56(1-3)*: 89-113 (2004)
4. S.Basu, Ian Davidson: Clustering with Constraints Theory and Practice. *ACM KDD 2006*
5. M.Badoiu, S.Har-Peled, P.Indyk, “Approximate clustering via core-sets”, *Proceedings of the 34th Symposium on Theory of Computing*, pp. 250–257, 2002.
6. C.Bhm, K.Kailing, P.Krger, A.Zimek, ”Computing Clusters of Correlation Connected Objects”.*Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD'04), Paris, France. pp. 455467. doi:10.1145/1007568.1007620.*
7. Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. *Journal of Computational Biology, 6(3/4):281-297, 1999.*
8. Ke Chen: On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications. *SIAM J. Comput. 39(3): 923-947 (2009)*
9. S. Dasgupta, ”The hardness of k-means clustering”. *Technical Report*, 2008.
10. Ayhan Demiriz, Kristin Bennett, and Mark J. Embrechts. Semi-supervised clustering using genetic algorithms. *In Artificial Neural Networks in Engineering, pages 809-814. ASME Press, 1999.*
11. Erik Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. ”Correlation clustering in general weighted graphs”. *Theor. Comput. Sci., 361(2):172-187*, 2006
12. Dan Feldman, Michael Langberg: A unified framework for approximating and clustering data. *STOC 2011: 569-578*
13. Dan Feldman, Morteza Monemizadeh, Christian Sohler: A PTAS for k-means clustering based on weak coresets. *Symposium on Computational Geometry 2007: 11-18*
14. H.Ding, J.Xu, ”Solving Chromatic Cone Clustering via Minimum Spanning Sphere”, *ICALP*, 2011
15. S. Har-Peled and S. Mazumdar, “Coresets for k-Means and k-Median Clustering and their Applications,” *Proc. 36th ACM Symposium on Theory of Computing, pages 291-300,* 2004.
16. Mary Inaba, Naoki Katoh, Hiroshi Imai, ” Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering (Extended Abstract)”. *Symposium on Computational Geometry 1994: 332-339*
17. S. G. Kolliopoulos and S. Rao, “A nearly linear-time approximation scheme for the euclidean k-median problem,” *Proc. 7th Annu. European Sympos. Algorithms, pages 378-389,* 1999.
18. A. Kumar, Y. Sabharwal, S. Sen, ” Linear-time approximation schemes for clustering problems in any dimensions”. *J. ACM 57(2):2010*
19. R.Ostrovsky, Y.Rabani, L.J.Schulman, and C.Swamy. ”The Effectiveness of Lloyd-Type Methods for the k-Means Problem”. *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). pp. 165174.*