Theoretical foundations for efficient clustering

by

Shrinu Kushagra

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Sanjoy Dasgupta Professor, Computer Science and Engineering University of California, San Diego
Supervisor(s):	Shai Ben-David Professor, Dept. of Computer Science, University of Waterloo
Internal Member:	Yaoliang Yu Assistant Professor, Dept. of Computer Science University of Waterloo
Internal-External Member:	Chaitanya Swamy Professor, Dept. of Combinatorics & Optimization University of Waterloo
Other Member(s):	Eric Blais Assistant Professor, Dept. of Computer Science University of Waterloo

Statement of Contributions

This thesis consists of material all of which I authored or co-authored. Chapter **??** is based on a joint work with Hassan Ashtiani and Shai Ben-David [**?**]. Chapter **??** is based on a joint work with Ihab Ilyas and Shai Ben-David. At the time of writing this thesis, the material covered in this chapter has not been published. But is going to appear in AISTATS'19 and ICDE'19. Chapter **??** is based on a joint work with Samira Samadi and Shai Ben-David [**?**]. Chapter **??** is based on a joint work with Yaoliang Yu and Shai Ben-David. At the time of writing this thesis, this work is still under submission at a conference. However, an arxiv version has been made available [**?**].

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

Abstract

Clustering is an umbrella term used to describe many common unsupervised learning techniques. One common view or definition of clustering is that it aims to group together data instances which are similar while simultaneously separating the dissimilar instances. Grouping objects into cohesive subsets is a fundamental problem in science and nature. We can find this problem in a variety of domains. Some examples include user segmentation, market analysis, dna sequencing, text analysis, image segmentation, data de-duplication and many more.

The task of clustering is challenging due to many factors. The most well-studied is the high computational cost. The clustering task can be viewed as an optimization problem where the goal is to minimize a certain cost function (like k-means cost or k-median cost). Not only are the minimization problems NP-Hard but often also NP-Hard to approximate (within a constant factor). There are two other major issues in clustering, namely *under-specificity* and *noise-robustness*. While some works have focussed on the issue of noise robustness of clustering algorithms, the problem of under-specificity has not received the adequate attention of the research community. The focus of this thesis is tackling these two issues while simultaneously ensuring low (polynomial) computational cost.

Clustering is under-specified. Consider the problem of dividing a dataset of human faces into two groups. One solution requirement could be to group the faces by gender. Another solution requirement could be to group the faces by emotion. Different solution requirements need different approaches. In such situations, domain knowledge is needed to better define the clustering problem. We incorporate this by allowing the clustering algorithm to interact with an oracle (or a human expert). The algorithm can ask the oracle whether two points belong to the same or different cluster. The oracle responds by replying either 'yes' or 'no' to the same-cluster query. In a preliminary work, we show that even access to a small number of same-cluster queries makes an otherwise NP-Hard clustering problem computationally tractable.

Pursuing this direction further, we consider the problem of clustering for data de-duplication; detecting records which correspond to the same physical entity in a database. Consider a marketing agency which sends advertising content (via emails, pamphlets, phone calls etc.) to potential consumers. The database of potential consumers is built from various sources and likely to contain duplicate records. In such cases, it is important to not the send the same content to a potential consumer multiple times. The framework of correlation clustering is highly applicable to model this problem. We propose a correlation clustering like framework to model such record de-duplication problems. We show that

access to a small number of same-cluster queries can help us solve the 'restricted' version of correlation clustering. Rather surprisingly, more relaxed versions of correlation clustering [1] are intractable even when allowed to make a 'large' (sub-linear in the size of the dataset) number of same-cluster queries.

The second line of research that this thesis explores is the issue of noise-robustness of clustering algorithms. Many of the common clustering tools (aim to) partition the data into cohesive groups. That is the groups or clusters share some between-cluster separation (the clustering community has introduced various notions of "clusterability" to capture this property). However, many real-world datasets, have on top of these cohesive subsets, a significant amount of points which are 'unstructured'. We refer to these structureless points as noise. The addition of these noisy points makes it difficult to detect the cohesive structure of the remaining points. The exact definition of 'structurelessness' varies depending on the type of structure the clustering algorithm is trying to detect.

In the first line of work, we define structurelessness as not having significantly large dense subsets. This definition is well suited to capture "gray background" noise contrasting with cohesive subsets of the data that the clustering aims to detect. We provide an computationally efficient clustering algorithm that captures all possible meaningful clusterings of the dataset (outputs a hierarchical clustering tree such that all meaningful clusterings are contained as a pruning in that tree). In this case, a meaningful solution is one where the clusters are cohesive (defined formally by notions of clusterability) and where the noise satisfies the gray background assumption.

Pursuing this line of research, we consider a second case where there is no restriction on the noisy points except that they are 'not too many'. That is the number of structureless points is small compared to the number of structured points. In this case, we develop a generic procedure that can transform any objective-based clustering algorithm into one that is robust to such noisy points. Our regularized transformation modifies any clustering objective function which outputs k clusters to one that outputs $k+1$ clusters. The algorithm is now allowed to discard a bunch of points into the extra garbage or noise cluster by paying a constant regularization penalty. Using this technique, we develop efficient noise-robust versions of two common algorithms. We show that both these algorithms are able to output a meaningful solution (under different assumptions on the clusterability of the cohesive subsets, of course).

[1] We refer to it as promise correlation clustering.

Acknowledgements

I would like to thank all the little people who made this thesis possible.

Dedication

This is dedicated to the one I love.

Table of Contents

List of Tables

List of Figures