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 [**?**]. The first part of Chapter **??** (the framework of promise correlation clustering) is based on a joint work with Ihab Ilyas and Shai Ben-David [**?**]. The experimental and hashing based algorithm in this chapter is based on the joint work with Hemant Saxena, Ihab Ilyas and Shai Ben-David [**?**]. Chapter **??** is based on a joint work with Yaoliang Yu and Shai Ben-David [**?**]. Chapter **??** is based on a joint work with Samira Samadi and Shai Ben-David [**?**]

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

Abstract

Clustering aims to group together data instances which are similar while simultaneously separating the dissimilar instances. 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*. The focus of this thesis is tackling these two issues while simultaneously ensuring low computational cost.

Clustering is an under-specified task. The same dataset may need to be clustered in different ways depending upon the intended application. 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 by asking whether two points belong to the same or different cluster. In a preliminary work, we show that access to a small number of same-cluster queries makes an otherwise NP-Hard k-means clustering problem computationally tractable. Next, we consider the problem of clustering for data de-duplication; detecting records which correspond to the same physical entity in a database. 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' number of same-cluster queries.

Next, we explore the issue of noise-robustness of clustering algorithms. Many real-world datasets, have on top of cohesive subsets, a significant amount of points which are 'unstructured'. The addition of these *noisy* points makes it difficult to detect the structure of the remaining points. In the first line of work, we develop a generic procedure that can transform objective-based clustering algorithms into one that is robust to outliers (as long the number of such points is not 'too large'). In particular, we develop efficient noise-robust versions of two common clustering algorithms and prove robustness guarantees for them. In the second line of work, we define noise as not having significantly large dense subsets. We provide computationally efficient clustering algorithms that capture all meaningful clusterings of the dataset; where the clusters are cohesive (defined formally by notions of clusterability) and where the noise satisfies the gray background assumption. We complement our results by showing that when either the notions of structure or the noise requirements are relaxed, no such results are possible.

[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