Example excerpted from *Sparse Methods for Learning Multiple Subspaces from Large-scale, Corrupted and Imbalanced Data* by Chong You, 2018.

Abstract

Not more than 350 words

In many practical applications in machine learning, computer vision, data mining and information retrieval one is confronted with datasets whose intrinsic dimension is much smaller than the dimension of the ambient space. This has given rise to the challenge of effectively learning multiple low-dimensional subspaces from such data. Multi-subspace learning methods based on sparse representation, such as sparse representation based classification (SRC) and sparse subspace clustering (SSC) have become very popular due to their conceptual simplicity and empirical success. However, there have been very limited theoretical explanations for the correctness of such approaches in the literature. Moreover, the applicability of existing algorithms to real world datasets is limited due to their high computational and memory complexity, sensitivity to data corruptions as well as sensitivity to imbalanced data distributions.

This thesis attempts to advance our theoretical understanding of sparse representation based multi-subspace learning methods, as well as develop new algorithms for handling large-scale, corrupted and imbalanced data. The first

First visible page number; title page is "i" but number is hidden
Front matter uses lower case roman numerals

contribution of this thesis is a theoretical analysis of the correctness of such methods. In our geometric and randomized analysis, we answer important theoretical questions such as the effect of subspace arrangement, data distribution, subspace dimension, data sampling density, and so on.

The second contribution of this thesis is the development of practical subspace clustering algorithms that are able to deal with large-scale, corrupted and imbalanced datasets. To deal with large-scale data, we study different approaches based on active support and divide-and-conquer ideas, and show that these approaches offer a good tradeoff between high accuracy and low running time. To deal with corrupted data, we construct a Markov chain whose stationary distribution can be used to separate between inliers and outliers. Finally, we propose an efficient exemplar selection and subspace clustering method that outperforms traditional methods on imbalanced data.

Include readers on the same pages as the abstract

Primary Reader and Advisor: René Vidal

Secondary Reader: Daniel P. Robinson

OPTIONAL

Acknowledgments

If included, place after abstract and before table of contents

I am very grateful to my advisor, Professor René Vidal, for introducing me to the fantastic world of machine learning, and for his guidance along the road of my PhD study. René's profound insights and broad vision in research have been a role model to me and have been the incentive to me to pursue my research. I am also thankful to him for always being patient and for giving me the freedom to explore different research ideas and conduct research through trial and error, which not only helped me to shape my independent research interests but also made the research process enjoyable.

I am also grateful to Professor Daniel P. Robinson for always being supportive in my research. I have learned a lot from Daniel on the importance of precise and rigorous writing in scientific papers and reports. I would also like to express my gratitude to Prof. Trac Tran, Prof. Vishal Patel, Prof. Enrique Mallada and Prof. Yi Ma for serving in my thesis proposal and dissertation committees, and to Prof. Gregory Hager, Prof. Sridevi Sarma for serving in my Graduate Board Oral committee.

OPTIONAL

Dedication

If included, place after the abstract and before the table of contents.

This thesis is dedicated to my parents, Fengyun Li and Zhendong You, for their eternal love, trust and support.

Acknowledgements, dedication, prefacing quotes, etc may all be placed in any order between the abstract and table of contents

Contents

Only required if tables are included in the document

List of Tables

If included, place after the table of contents and before the list of figures

Includes table number and accurate page number

Still part of front matter; receives lower case roman numer page number

Only required if figures are included in the document

List of Figures

If included, place after list of tables

Includes figure number and accurate page number

Still part of front matter; receives lower case roman numer page number

Marks beginning of Main Text

Chapter 1

Introduction

The significant increase in the ability to collect and store diverse information in the past decades has led to an exceptional growth in the availability of data. In the field of computer vision, for instance, portable and affordable digital cameras and smartphones interconnected with high-speed mobile networks have produced image and video datasets of unprecedented scale, which are being collected by giant Internet companies such as Google and Amazon through services they provide to billions of customers. The proliferation in dataset size and complexity is accompanied by the challenge of successfully analyzing the data to discover patterns of interest. Aside from being large-scale, modern datasets very often possess significant amounts of corruptions in various forms such as noise, corrupted entries, outliers and missing entries. All these features pose stark challenges to the development of techniques for modern data

Arabic numeral
Re-start page numbering here