UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

DISTANCE-BASED DIMENSIONALITY REDUCTION FOR BIG DATA

ADRIÀ CASANOVA LLOVERAS

Thesis supervisor
PEDRO DELICADO USEROS (Department of Statistics and Operations Research)

Thesis co-supervisor
CRISTIAN PACHÓN GARCIA (Department of Statistics and Operations Research)

Degree
Master's Degree in Data Science

Master's thesis

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Abstract

Dimensionality reduction aims to project a data set into a low-dimensional space. Many techniques have been proposed, most of them based on the inter-individual distance matrix. When the number of individuals is really large, the use of distance matrices is prohibitive. There are algorithms that extend MDS (a classical dimensionality reduction method based on distances) to the big data setting. In this TFM, we adapt these algorithms to any generic distance-based dimensionality reduction method.

Contents

1 Introduction, Motivation, and Objectives

1.1 Introduction

- Dimensionality Reduction definition, goal and applications.
- Examples of DR methods.
- Key points and limitations of DR methods.

1.2 Motivation

- When the number of individuals is really large, the use of distance matrices is prohibitive.
- There are algorithms that extend MDS to the big data setting.

1.3 Objectives

- Adapt these algorithms to any generic distance-based dimensionality reduction method.

2 State of the Art

- R libraries: `Rdimtools`, `dimRed`.
- `bigmds` [1].
- Landmark Isomap [3].
- Out-of-Core Dimensionality Reduction for Large Data via Out-of-Sample Extensions [2].
- t-SNE Python implementations.

3 Specification and Design of the Solution

- D&C algorithm

4 Development of the Proposal

- DR methods descriptions.
- DR methods implementations.

5 Experimentation and Evaluation of the Proposal

- Swiss Roll.
- MNIST.
- Tetrahedron.
- Time complexity measurements (including MDS).
- Evaluation and comparison with state of the art.

6 Analysis of Sustainability and Ethical Implications

It must include an analysis of the impact of the following gender-related technical aspects:

- issues related to data management and analysis
- issues related to equity, where possible biases are identified and assessed both in the data and in the processes carried out in relation to data management and analysis
- actions carried out to eliminate or mitigate such biases

7 Conclusions

References

[1] Pedro Delicado and Carlos Pachón-García. "Multidimensional scaling for big data". In: *Advances in Data Analysis and Classification* (2024). ISSN: 1862-5355. DOI: 10.1007/s11634-024-00591-9.

[2] Luca Reichmann, David Hägele, and Daniel Weiskopf. *Out-of-Core Dimensionality Reduction for Large Data via Out-of-Sample Extensions*. 2024. arXiv: 2408.04129 [cs.LG]. URL: https://arxiv.org/abs/2408.04129.

[3] Vin Silva and Joshua Tenenbaum. "Global Versus Local Methods in Nonlinear Dimensionality Reduction". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/5d6646aad9bcc0be55b2c82f69750387-Paper.pdf.

[4] Kisung You and Dennis Shung. "Rdimtools: An R Package for Dimension Reduction and Intrinsic Dimension Estimation". In: *Software Impacts* 14 (2022), p. 100414. ISSN: 26659638. DOI: 10.1016/j.simpa.2022.100414.

8 Annexes

Example citation: [4].