



Universiteit  
Leiden

The Netherlands

## Exploring graph-based clustering and outlier detection algorithms

Li, J.

### Citation

Li, J. (2025, November 12). *Exploring graph-based clustering and outlier detection algorithms*. Retrieved from <https://hdl.handle.net/1887/4282945>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4282945>

**Note:** To cite this publication please use the final published version (if applicable).

# Exploring graph-based clustering and outlier detection algorithms

Proefschrift

ter verkrijging van  
de graad van doctor aan de Universiteit Leiden,  
op gezag van rector magnificus prof.dr.ir. H. Bijl,  
volgens besluit van het college voor promoties  
te verdedigen op woensdag 12 november 2025  
klokke 10:00 uur

door

Jia Li  
geboren te Henan, China  
in 1992

**Promotor:**

Prof.dr.ir. F.J. Verbeek

**Co-promotor:**

Dr. C. Wang (Xi'an Jiaotong University)

**Promotiecommissie:**

Prof.dr. M.M. Bonsangue

Prof.dr. K.J. Batenburg

Prof.dr. A. Plaat

Prof.dr. B. Nurami Ruchjana, Universitas Padjadjaran, Bandung Indonesia

Dr. Y. Guo, Research Institute of Highway, China

Dr. R. Pucci

Copyright © 2023 Jia Li

ISBN:

The research is financially supported by the Chinese Scholarship Council  
(CSC 202106280008)

“Nothing can be accomplished without norms or standards.”

Mencius, 372-289 BCE

# Contents

|          |                                                                                                            |           |
|----------|------------------------------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                                                        | <b>1</b>  |
| 1.1      | Research Background . . . . .                                                                              | 1         |
| 1.2      | Machine Learning Algorithms . . . . .                                                                      | 4         |
| 1.2.1    | Minimum Spanning Tree-based clustering algorithms . . . . .                                                | 4         |
| 1.2.2    | Spectral clustering algorithms . . . . .                                                                   | 5         |
| 1.2.3    | Clustering-based outlier detection algorithms . . . . .                                                    | 6         |
| 1.3      | Research Questions . . . . .                                                                               | 7         |
| 1.4      | Structure of This Thesis . . . . .                                                                         | 8         |
| <b>2</b> | <b>A Scaled Minimum Spanning Tree Based Clustering Algorithm<br/>and Application on Image Segmentation</b> | <b>11</b> |
| 2.1      | Introduction . . . . .                                                                                     | 13        |
| 2.2      | Related work . . . . .                                                                                     | 15        |
| 2.2.1    | MST-based clustering algorithms . . . . .                                                                  | 15        |
| 2.2.2    | Image segmentation and integration . . . . .                                                               | 17        |
| 2.3      | The proposed method . . . . .                                                                              | 18        |
| 2.3.1    | Scaled-MST . . . . .                                                                                       | 19        |
| 2.3.2    | Scaled-MST-based clustering . . . . .                                                                      | 20        |
| 2.3.3    | Our second scaled-MST-based clustering algorithm . . . . .                                                 | 23        |
| 2.3.4    | Image segmentation and integration . . . . .                                                               | 26        |
| 2.3.5    | Time complexity analysis . . . . .                                                                         | 27        |
| 2.4      | Evaluation . . . . .                                                                                       | 28        |
| 2.4.1    | Results on synthetic data sets . . . . .                                                                   | 28        |
| 2.4.2    | Results on real data sets . . . . .                                                                        | 31        |
| 2.4.3    | Results on image . . . . .                                                                                 | 34        |

|          |                                                                                        |           |
|----------|----------------------------------------------------------------------------------------|-----------|
| 2.5      | Conclusion . . . . .                                                                   | 42        |
| <b>3</b> | <b>A Novel Approximate MST-based Clustering Algorithm based on Sampling</b>            | <b>45</b> |
| 3.1      | Introduction . . . . .                                                                 | 47        |
| 3.2      | Related work . . . . .                                                                 | 49        |
| 3.2.1    | MST algorithm . . . . .                                                                | 49        |
| 3.2.2    | MST-based clustering algorithm . . . . .                                               | 50        |
| 3.3      | Our method . . . . .                                                                   | 51        |
| 3.3.1    | Basic idea . . . . .                                                                   | 52        |
| 3.3.2    | Proposed SMST method . . . . .                                                         | 53        |
| 3.3.3    | Our SMST-based clustering method . . . . .                                             | 55        |
| 3.3.4    | Complexity analysis . . . . .                                                          | 57        |
| 3.4      | Evaluation . . . . .                                                                   | 58        |
| 3.4.1    | Results on synthetic data sets . . . . .                                               | 58        |
| 3.4.2    | Results on real-world data sets . . . . .                                              | 63        |
| 3.4.3    | Sampling size analysis . . . . .                                                       | 65        |
| 3.4.4    | Running time performance . . . . .                                                     | 67        |
| 3.5      | Conclusion . . . . .                                                                   | 69        |
| <b>4</b> | <b>Outlier Detection in Eigenvalue Spaces based on Spectral Analysis of KNN Graphs</b> | <b>71</b> |
| 4.1      | Introduction . . . . .                                                                 | 73        |
| 4.2      | Related work . . . . .                                                                 | 75        |
| 4.2.1    | Density-based outlier detection . . . . .                                              | 75        |
| 4.2.2    | Space transformation based outlier detection . . . . .                                 | 77        |
| 4.2.3    | Multi-view and deep learning-based outlier detection . . . . .                         | 78        |
| 4.3      | Spectral analysis based outlier detection . . . . .                                    | 79        |
| 4.3.1    | Background to our method . . . . .                                                     | 79        |
| 4.3.2    | Spectral Analysis . . . . .                                                            | 81        |
| 4.3.3    | Outlier detection . . . . .                                                            | 83        |
| 4.3.4    | Time complexity analysis . . . . .                                                     | 87        |
| 4.4      | Experiments . . . . .                                                                  | 87        |
| 4.4.1    | Experimental setup . . . . .                                                           | 88        |
| 4.4.2    | Results . . . . .                                                                      | 90        |

|          |                                                                                                                           |            |
|----------|---------------------------------------------------------------------------------------------------------------------------|------------|
| 4.4.3    | Sensitivity to the parameter $k$ . . . . .                                                                                | 92         |
| 4.4.4    | Eigenvalue distribution . . . . .                                                                                         | 96         |
| 4.5      | Conclusions . . . . .                                                                                                     | 97         |
| <b>5</b> | <b>Outlier Detection Using Iterative Adaptive Mini-Minimum Spanning Tree Generation with Applications on Medical Data</b> | <b>99</b>  |
| 5.1      | Introduction . . . . .                                                                                                    | 101        |
| 5.2      | Related work . . . . .                                                                                                    | 103        |
| 5.2.1    | Distance-based outlier detection . . . . .                                                                                | 103        |
| 5.2.2    | Minimum spanning tree-based outlier detection . . . . .                                                                   | 104        |
| 5.2.3    | Summary of deficiencies . . . . .                                                                                         | 104        |
| 5.3      | Foundation . . . . .                                                                                                      | 105        |
| 5.3.1    | Preliminaries . . . . .                                                                                                   | 105        |
| 5.3.2    | Definitions . . . . .                                                                                                     | 105        |
| 5.4      | Method . . . . .                                                                                                          | 107        |
| 5.4.1    | MST generation details and an illustrative example . . . . .                                                              | 107        |
| 5.4.2    | Adaptive mini-minimum spanning tree-based outlier detection (MMOD) . . . . .                                              | 108        |
| 5.5      | Experimental Results and Evaluation . . . . .                                                                             | 112        |
| 5.5.1    | Applied datasets . . . . .                                                                                                | 112        |
| 5.5.2    | State-of-the-art methods for comparison . . . . .                                                                         | 112        |
| 5.5.3    | Evaluation metrics . . . . .                                                                                              | 113        |
| 5.5.4    | Results on real-world datasets . . . . .                                                                                  | 114        |
| 5.5.5    | Comprehensive performance analysis and discussion . . . . .                                                               | 119        |
| 5.6      | Conclusion . . . . .                                                                                                      | 120        |
| <b>6</b> | <b>MS2OD: Outlier Detection Using Minimum Spanning Tree and Medoid Selection</b>                                          | <b>131</b> |
| 6.1      | Introduction . . . . .                                                                                                    | 133        |
| 6.2      | Existing methods and their limitations . . . . .                                                                          | 135        |
| 6.2.1    | Minimum spanning tree (MST) . . . . .                                                                                     | 135        |
| 6.2.2    | Centroid selection . . . . .                                                                                              | 136        |
| 6.3      | Proposed method . . . . .                                                                                                 | 137        |
| 6.3.1    | Ideas for methodological improvements . . . . .                                                                           | 137        |
| 6.3.2    | Medoid selection and outlierness degree . . . . .                                                                         | 137        |

**Contents**

---

|          |                                                              |            |
|----------|--------------------------------------------------------------|------------|
| 6.3.3    | Outlier detection using MST and medoid selection (MS2OD)     | 139        |
| 6.4      | Experimental setup and results . . . . .                     | 140        |
| 6.4.1    | Peer methods for comparison and experimental configurations  | 140        |
| 6.4.2    | Illustrative comparison on a typical synthetic dataset . . . | 141        |
| 6.4.3    | Real-world datasets . . . . .                                | 141        |
| 6.4.4    | Metrics . . . . .                                            | 142        |
| 6.4.5    | Results . . . . .                                            | 143        |
| 6.5      | Evaluation and Discussion . . . . .                          | 143        |
| 6.5.1    | Outlier Detection DataSets (ODDS) . . . . .                  | 143        |
| 6.5.2    | DAMI’s “datasets used in the literature” category . . .      | 144        |
| 6.5.3    | DAMI’s “semantically meaningful datasets” category . . .     | 146        |
| 6.5.4    | Comprehensive Analysis . . . . .                             | 147        |
| 6.6      | Conclusion . . . . .                                         | 148        |
| <b>7</b> | <b>Conclusion and Outlook</b>                                | <b>155</b> |
| 7.1      | Conclusions . . . . .                                        | 156        |
| 7.2      | Discussion . . . . .                                         | 159        |
| 7.3      | Future Research Directions . . . . .                         | 160        |
|          | <b>References</b>                                            | <b>163</b> |
|          | <b>Samenvatting</b>                                          |            |
|          | <b>Summary</b>                                               |            |
|          | <b>摘要</b>                                                    |            |
|          | <b>Acknowledgements</b>                                      |            |
|          | <b>Curriculum Vitae</b>                                      |            |
|          | <b>Publications</b>                                          |            |