

Advanced Clustering

by Jie Yang

Thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy

under the supervision of

Principal Supervisor: Prof. Chin-Teng Lin

Co-Supervisor: Dr. Yu-Kai Wang

University of Technology Sydney

Faculty of Engineering and Information Technology

May 2022

Certificate of Original Authorship

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Jie Yang*, declare that this thesis is submitted in fulfilment of the requirements for the award of the *Doctoral Degree*, in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

**If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).*

**If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).*

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 2/April/2023

Collaborative doctoral research degree statement

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with *[insert collaborative partner institution]*.

Indigenous Cultural and Intellectual Property (ICIP) statement

This thesis includes Indigenous Cultural and Intellectual Property (ICIP) belonging to *[insert relevant language, tribal or nation group(s) or communities]*, custodians or traditional owners. Where I have used ICIP, I have followed the relevant protocols and consulted with appropriate Indigenous people/communities about its inclusion in my thesis. ICIP rights are Indigenous heritage and will always remain with these groups. To use, adapt or reference the ICIP contained in this work, you will need to consult with the relevant Indigenous groups and follow cultural protocols.

Acknowledgement

I would like to express my heartfelt appreciation to my principal supervisor, Professor CT Lin, for his unwavering guidance, encouragement, and support throughout my Ph.D. journey. I am grateful for his generosity in allowing me to pursue research topics that interest me. During my time with Professor CT Lin, I have gained valuable insights and learned a great deal. I would also like to extend my thanks to my co-supervisor, Dr. YK Wang, for his invaluable assistance, guidance, and mentorship throughout my Ph.D. studies.

My deepest gratitude goes to the CIBCI Lab and its members, who have provided me with various resources and selfless support during my research. I would like to extend a special thank you to Xiaofei Wang, Jia Liu, Fred Chang, and Liang Ou for their help and support in overcoming the challenges that I faced.

I also want to thank my friends at TechLab for the fun and memorable times we shared together, and for their continued support in my academic journey. Furthermore, I am grateful to all the friends I have made at UTS, who have provided various forms of assistance and support.

I would like to acknowledge the financial support of the Australian Research Council (ARC) under discovery grants DP180100656 and DP210101093. Additionally, I would like to express my gratitude to the UTS International Research Scholarship for covering my tuition fees.

My heartfelt appreciation goes to my parents, who have been a constant source of love, support, and guidance. I am grateful for their unwavering encouragement and understanding.

Finally, I would like to thank myself for my hard work and persistence, especially during the challenging times caused by the COVID-19 pandemic. Despite the difficult circumstances, I remained committed to my research and was able to complete this project.

Published and Under Review Papers Related to This Thesis

- [1] **J. Yang** and C.-T. Lin, “Multi-View Adjacency-Constrained Hierarchical Clustering”, *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol. Early Access, pp. 1-13, 2022 [**Chapter 4**]
- [2] **J. Yang**, Y.-K. Wang, X. Yao, and C.-T. Lin, “Adaptive Initialization Method for K-Means Algorithm,” *Frontiers in Artificial Intelligence*, vol. 4, 2021 [**Chapters 1-2**]
- [3] **J. Yang** and C.-T. Lin, “Multi-View Adjacency-Constrained Nearest Neighbor Clustering (Student Abstract),” *AAAI-2022*, Vol. 36, No. 11, pp. 13097-13098, 2022 [**Chapter 4**]
- [4] **J. Yang** and C.-T. Lin, “Autonomous clustering by fast find of mass and distance peaks”, submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Major Revision) [**Chapter 3**]
- [5] **J. Yang** and C.-T. Lin, “PSO-based Multi-View Nearest Neighbor Clustering”, submitted to *IEEE Computational Intelligence Magazine* (Under Review) [**Chapter 5**]
- [6] **J. Yang** and C.-T. Lin, “Enhanced Adjacency-constrained Hierarchical Clustering using Fine-grained Pseudo labels”, submitted to *IEEE Transactions on Emerging Topics in Computational Intelligence* (Under Review)
- [7] **J. Yang** and C.-T. Lin, “Almost Ultrametric Learning using Pseudo Labels from Clustering” (Draft) [**Chapter 6**]
- [8] **J. Yang** and C.-T. Lin, “Improve Torque Clustering by optimizing linkage” (Draft)
- [9] **J. Yang** and C.-T. Lin, “Distributed Torque Clustering” (Draft)

Abstract

Clustering is a classical technique in the field of data mining. It has played a key role in domains such as biology, medicine, business, and climatology, and is employed in nearly all scientific and social sciences. Despite the significance and pervasiveness of clustering and the plethora of existing algorithms, the current clustering methods suffer from a variety of drawbacks. For example, standard hierarchical clustering has an excessive computational overhead and requires some manually determined conditions. Partition clustering, such as K-means, demands that the number of clusters must either be known or estimated in advance and cannot detect non-convex clusters of varying size or density. Density clustering typically requires a suite of thresholds to be set in advance, such as cut-off distance. Model-based clustering generally relies on prior knowledge of many parameter settings, which is often very difficult to acquire in practice. Classic grid clustering also depends on many user-provided parameters, such as interval values to divide space and density thresholds.

On the other hand, in recent years, multi-view clustering has become a new research hotspot. Essentially, multi-view clustering arises from the combination of clustering problems and multi-view learning. Different from the various conventional single-view clustering methods mentioned above, as an extension of single-view clustering, multi-view clustering is used to handle multi-view data gathered from numerous feature collectors or collected from various sources in various domains. However, most current multi-view clustering approaches suffer from the following three problems: a) parameter tuning, b) significant computational cost, and c) difficulty in finding globally optimal view weights.

To solve the above problems, this thesis first proposes a brand-new efficient parameter-free autonomous clustering algorithm called Torque Clustering (TC). The proposed TC overcomes almost all the shortcomings in previous clustering methods. Furthermore, considering the good performance of the proposed TC, this thesis extends TC to two multi-view clustering algorithms, containing multi-view adjacency-constrained hierarchical clustering (MCHC) and particle swarm optimization (PSO)-based multi-view nearest neighbor clustering (PMNNC). MCHC tries to solve two problems in current multi-view clustering methods: a) parameter tuning and b)

significant computational cost. PMNNC focuses on solving the third problem: c) difficulty in finding globally optimal view weights. Finally, we further apply the pseudo labels generated by TC to propose a new metric learning framework, named almost ultrametric learning using pseudo labels of torque clustering (AUMLTC), which can help other algorithms improve performance in a parameter-free and unsupervised manner.

This Ph.D. thesis contains seven chapters. Chapter 1 introduces the background, objectives, scope, organization, and contributions of the thesis. Chapter 2 presents the literature review of the research. Chapter 3 proposes a new parameter-free autonomous clustering, i.e., TC. Chapter 4 exploits the partial mechanism of TC in Chapter 3 as a backbone to propose a new parameter-free multi-view clustering with low computational overhead, i.e., MCHC. Chapter 5 also exploits the partial mechanism of TC in Chapter 3 as a backbone to propose a novel multi-view clustering based on an evolutionary algorithm, i.e., PMNNC. Chapter 6 leverages the pseudo labels of TC in Chapter 3 to propose a new metric learning framework, i.e., AUMLTC. Chapter 7 includes an overview of the thesis's contents and some suggestions for future works.

Keywords: Clustering, Parameter-free, Multi-view Clustering, Autonomous, Metric Learning

Contents

Abstract.....	V
Chapter 1. Introduction.....	1
1.1 Background.....	1
1.2 Insights and our solutions.....	4
1.3 Research objectives.....	5
1.4 Research scope.....	6
1.5 Thesis overview.....	7
1.6 Key contributions.....	8
Chapter 2. Literature review.....	11
2.1 Clustering.....	11
2.1.1 Hierarchical clustering.....	11
2.1.2 Partition-based clustering.....	14
2.1.3 Density-based clustering.....	16
2.1.4 Model-based clustering.....	18
2.1.5 Grid-based clustering.....	19
2.2 Multi-view clustering.....	21
2.2.1 Multi-view spectral clustering.....	21
2.2.2 Multi-view subspace clustering.....	22
2.2.3 Other multi-view clustering.....	22
2.3 Distance metric learning.....	23
2.3.1 Supervised metric learning.....	24
2.3.2 Semi-supervised metric learning.....	24
2.3.3 Unsupervised metric learning.....	25
2.4 Summary.....	26
Chapter 3. Torque Clustering: Autonomous clustering by fast find of mass and distance peaks ...	27
3.1 Introduction.....	27
3.2 Proposed method.....	29
3.2.1 Define clusters and form connections between them.....	29
3.2.2 Define two properties of each connection to construct the decision graph.....	30
3.2.3 Define the torque of each connection and sort the connections in descending order.....	34
3.2.4 Define torque gap and find the largest gap to determine abnormal connections.....	35
3.2.5 Define halo connections to determine the noise.....	36
3.2.6 Complexity analysis.....	37
3.2.7 Algorithm analysis.....	39
3.3 Experiments and results.....	39
3.3.1 Evaluation on nine synthetic data sets.....	40
3.3.2 Evaluation on 11 real-world data sets.....	42
3.3.3 Results and analysis.....	45
3.3.4 Runtime.....	48
3.3.5 Further evaluation on 56 data sets with peculiar characteristics.....	49
3.3.6 Comparison to deep clustering algorithms on challenging image data sets.....	49
3.4 Discussion.....	50

3.4.1 Differences between TC and other hierarchical clustering algorithms	50
3.4.2 Differences between TC and density peak clustering algorithms	51
3.4.3 Differences between TC and subspace clustering algorithms	52
3.4.4 Potential limitations of TC	52
3.5 Conclusion	52
3.6 Experimental details and more results	54
3.6.1 Experimental details	58
3.6.2 Further evaluation on additional 56 data sets with peculiar characteristics	63
3.6.3 Robustness guarantees for the proposed TC	71
Chapter 4. Multi-view adjacency-constrained hierarchical clustering	74
4.1 Introduction	74
4.2 Proposed method	76
4.2.1 Fusion distance matrices with extreme weights (FDEW)	76
4.2.2 Adjacency-constrained nearest neighbor clustering (CNNC)	78
4.2.3 Internal evaluation index based on Rawls' max-min criterion (MMI)	81
4.2.4 Algorithm of MCHC and MCHC-PF	82
4.3 Experiments and results	86
4.3.1 Data sets description	86
4.3.2 Compared algorithms	87
4.3.3 Results and analysis	88
4.3.4 Runtime	90
4.4 Ablation study	91
4.4.1 Impact of fusion distance matrices with extreme weights (FDEW)	91
4.4.2 Impact of adjacency-constrained nearest neighbor clustering (CNNC)	92
4.4.3 Impact of internal evaluation index based on Rawls' max-min criterion (MMI)	94
4.5 Discussion	94
4.6 Conclusion	96
Chapter 5. PSO-based multi-view nearest neighbor clustering	97
5.1 Introduction	97
5.2 Proposed Method	98
5.2.1 Adjacency-constrained nearest neighbor clustering (CNNC)	98
5.2.2 Particle swarm optimization (PSO)	100
5.2.3 Fitness function based on a novel internal validity index: minimum spanning tree-based Dunn's index (MSTDI)	101
5.2.4 PSO-based multi-view nearest neighbor clustering	102
5.3 Experiment and results	104
5.3.1 Data sets description	104
5.3.2 Compared algorithms	105
5.3.3 Results and analysis	106
5.4 Ablation study	109
5.4.1 Impact of adjacency-constrained nearest neighbor clustering (CNNC)	109
5.4.2 Impact of minimum spanning tree-based Dunn's index (MSTDI)	109
5.4.3 Impact of the two hyperparameters	110
5.5 Conclusion	111
Chapter 6. Almost ultrametric learning using pseudo labels from clustering	112

6.1 Introduction	112
6.2 Proposed Method	113
6.2.1 Metric space and ultrametric space	113
6.2.2 Convert metric space to the proposed almost ultrametric (AUM) space	114
6.2.3 Exploit pseudo labels of torque clustering to approximate ground truth labels	116
6.2.4 Almost ultrametric learning using TC pseudo labels (AUMLTC)	119
6.3 Experiments and results	120
6.3.1 Data sets description	120
6.3.2 Compared algorithms	121
6.3.3 Results and analysis	121
6.4 Ablation study	123
6.5 Conclusion	124
Chapter 7. Conclusion and future work	127
7.1 Conclusion	127
7.2 Future work	128
Bibliography	130