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.

Signature: Production Note: 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