University of Nevada
Reno

Robust Fuzzy Cluster Ensemble on Cancer Gene Expression Data

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering

by

Yan Yan

Dr. Frederick C. Harris, Jr./Dissertation Advisor

May, 2019

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

YAN YAN

Entitled

Robust Improved Fuzzy Cluster Ensemble On Cancer Gene Expression Data

be accepted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Dr. Frederick C. Harris, Jr., Advisor

Dr. Sergiu Dascalu, Committee Member

Dr. Dwight Egbert, Committee Member

Dr. Ania Panorska, Committee Member

Dr. Tin Nguyen, Committee Member

Dr. Yantao Shen, Graduate School Representative

David W. Zeh, Ph. D., Dean, Graduate School
May, 2019

Abstract

In the past few decades, there has been tremendous growth in the scale and complexity of biological data generated by emerging high-throughput biotechnologies, including gene expression data generated by microarray technology. High-throughput gene expression data may contain gene expression measurements of thousands or millions of genes in a single data set, and provide us opportunities to explore the cell on a genome wide scale. Finding patterns in genomic data is a very important task in bioinformatics research and biomedical applications. Many clustering algorithms have been applied to gene expression data to find patterns. Nonetheless, there are still a number of challenges for clustering gene expression data because of the specific characteristics of such data and the special requirements from the domain of biology. Data noise and data high dimensionality are among the top challenges.

In this dissertation, we propose a novel fuzzy cluster ensemble methodology which is effective and efficient in addressing the data noise and data high dimensionality challenges. It consists of an improved fuzzy clustering approach with different initializations as its base clusterings in order to reduce the impact of noises and improve accuracy and stability in general. The improved fuzzy clustering approach uses new weighted fuzzy techniques in computing cluster centers and assigning feature vectors, to avoid or alleviate the effects of noise.

We conducted extensive experiments for our methodology on both real cancer gene expression data sets and synthetic noisy data sets created by introducing different percentages of artificial noise to real cancer gene expression data sets. We chose an external clustering validity measure for evaluating domain meaningfulness. For experiments on real cancer gene expression data sets, the results were evaluated using comparisons with numerous benchmark clustering and cluster ensemble algorithms. We also conducted parameter analysis on various parameters with different settings, complexity analysis on time cost and space cost, and noise robustness analysis on synthetic noisy data sets. The results from real cancer gene expression data sets have proved to be biologically and medically meaningful. Our methodology is the top

performer on three of the eight data sets, more than any other methods evaluated, and it performs well on most of the other data sets. Additionally, our methodology have proved to be stable with varying parameter settings. For complexity analysis on time cost and space cost, it is computational efficient and scalable to high dimensional data sets. For noise robustness analysis experiments, the results have proved to be robust against highly noisy data.

Dedication

For my beloved family and all the people who have inspired me and supported me.

Acknowledgments

I would like to express great appreciation to all the people who supported and helped me tremendously during my study.

First of all, I would like to thank my advisor Dr. Frederick C. Harris, Jr. Over the years, I have been very grateful for having his continuous knowledgeable and experienced guidance, as well as his kind support with patience and a warm heart. He has been so encouraging and helpful in so many ways, especially during challenging times. I am truly feeling grateful to have such a supportive advisor. My special thanks to Ms. Cindy Harris for meticulously reading and editing a very long paper of mine. She is so kind and really generous.

I would like to thank my committee members, Dr. Sergiu Dascalu, Dr. Dwight Egbert, Dr. Ania Panorska, Dr. Yantao Shen, and Dr. Tin Nguyen, for their valuable time and invaluable feedback not only on my proposal and dissertation, but also on research papers. I would like to thank my M.S. Thesis advisor Dr. Carl Looney for his knowledgeable and experienced guidance as well as kind support till his retirement. They are among the most helpful and supportive.

I would like to thank other faculty members and staff in the Computer Science and Engineering Department for being always helpful and promptly supportive in various areas during the course of my studies.

I would like to thank my family for all their unconditional love, encouragement, and support in my pursuits during the long journey. I would also like to thank my friends for their generous support and help along the way.

Contents

List of Figures

List of Tables