

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

Copyright by Yan Yan 2018

All Rights Reserved



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

Abstract	i
List of Figures	vi
List of Tables	ix
1 Introduction	1
1.1 Clustering and Bioinformatics	1
1.2 Motivation	6
1.3 Methodology	8
1.4 Contributions	8
1.5 Dissertation Organization	9
2 Background and Literature Review	10
2.1 Introduction	10
2.2 Data Clustering	13
2.2.1 Hierarchical Clustering	14
2.2.2 Partitioning Clustering	17
2.2.3 Graph-based Clustering	19
2.2.4 Distribution-based Clustering	20
2.2.5 Density-based Clustering	21
2.2.6 Grid-based Clustering	23
2.2.7 Clustering High Dimensional Data	24
2.2.8 Other Clustering Techniques	27
2.3 Applications of Clustering in Cancer Subtyping	35
2.3.1 Clinical Applications	37
2.3.2 Computational Experiments	41
2.4 Challenges	44
2.4.1 Clinical Challenges	44
2.4.2 Computational Challenges	45
2.5 Discussion	46
3 Improved Fuzzy Cluster Ensemble Methodology	49
3.1 Introduction	49

3.2	Related Work	50
3.3	Noise Robustness Problem	51
3.4	Fuzzy Set Theory	53
3.5	Improved Fuzzy Clustering Algorithm	57
3.5.1	Description	57
3.5.2	High Level Algorithm and Flowchart	62
3.5.3	Iris Data Set	64
3.6	Improved Fuzzy Cluster Ensemble Algorithm	67
3.6.1	Diagram	67
3.6.2	Ensemble Generation	68
3.6.3	Ensemble Consensus	69
4	Experimental Results	70
4.1	Experiment Design and Settings	70
4.1.1	Cancer Gene Expression Data Sets	70
4.1.2	Comparable Clustering Algorithms	73
4.1.3	Validity Measure	74
4.1.4	Number of Clusters	77
4.2	Validity Measure Comparison	77
4.3	Parameter Analysis	82
4.3.1	N (number of clustering runs)	82
4.3.2	IMT (initial merging threshold)	83
4.3.3	M (ensemble size)	85
4.4	Complexity Analysis	87
4.4.1	Time Complexity	87
4.4.2	Space Complexity	88
4.5	Noise Robustness Analysis	88
4.5.1	Synthetic Noisy Data Sets	88
4.5.2	Results	88
4.6	Conclusion	89
5	Conclusions and Future Work	91
5.1	Conclusions	91
5.2	Future Work	93
5.2.1	Clustering Algorithm	94
5.2.2	Bioinformatics	94
	Bibliography	97

List of Figures

1.1	Bioinformatics and related disciplines [267]	2
1.2	Bioinformatics applications [191]	2
1.3	Clustering in bioinformatics [184]	3
2.1	An example of hierarchical clustering applied on gene expression data. In this example, the rows represent the genes while the columns represent different samples. The expression values are color coded (from red to green). The hierarchical clustering are performed on both rows (genes) and columns (samples).	16
2.2	An example of k-means based clustering [208] on a lung cancer dataset [122]. The data shown in the space of the first three principal components. Different colors represent different clusters.	19
3.1	Data and signal [256]	51
3.2	Data and noise [255]	52
3.3	Fuzzy logic [17]	54
3.4	Fuzzy logic examples [245]	54
3.5	Fuzzy membership [224]	55
3.6	Parameterized fuzzy membership functions [229]	57
3.7	Sigmoid fuzzy membership functions [229]	58
3.8	Left-Right(L-R) fuzzy membership functions [229]	58
3.9	Gaussian fuzzy set membership function [3]	59
3.10	Modified weighted fuzzy expected value [174]	60
3.11	An example of cluster merging [137]	62
3.12	Flowchart of IFC	65
3.13	Petal and sepal of Iris flower [268]	66
3.14	Three species of Iris flower [213]	66
3.15	Spectramap biplot of Fisher's Iris data set [192]	67
3.16	MWFEV centers of the four Iris features [174]	67
3.17	IFCE (adapted from [264])	68
4.1	Common external validity measures [242]	75

4.2	Common internal validity measures [242]	75
4.3	Common relative validity measures [242]	76
4.4	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM across eight real cancer gene expression data sets over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	78
4.5	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Golub1999v1 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	79
4.6	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Golub1999v2 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	79
4.7	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Armstrong2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	79
4.8	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Chowdary2006 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	80
4.9	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Nutt2003 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	80
4.10	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Pomeroy2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	80
4.11	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Chen2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	81
4.12	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Khan2001 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	81

4.13	CA of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.	83
4.14	Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.	84
4.15	CA of IFCE on Chowdary2006 and Chen2002 with IMT (initial merging threshold) = 1.0, 2.0, 3.0, 4.0.	85
4.16	Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with IMT (initial merging threshold) = 1.0, 2.0, 3.0, 4.0.	85
4.17	CA of IFCE on Chowdary2006 and Chen2002 with M (ensemble size) = 3, 7, 11, 21.	86
4.18	Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with M (ensemble size) = 3, 7, 11, 21.	87
4.19	Noise Robustness with artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% added to Chowdary2006 and Chen2002.	90

List of Tables

3.1	Fuzzy membership functions	56
4.1	Cancer gene expression data sets	73
4.2	Comparable clustering algorithms	74
4.3	Comparable cluster ensemble algorithms	74
4.4	CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM across eight real cancer gene expression data sets over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].	77
4.5	CA of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.	82
4.6	Run Time (sec.) of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.	83
4.7	CA of IFCE on Chowdary2006 and Chen2002 with IMT(initial merging threshold) = 1.0, 2.0, 3.0, 4.0.	84
4.8	Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with IMT(initial merging threshold) = 1.0, 2.0, 3.0, 4.0.	84
4.9	CA of IFCE on Chowdary2006 and Chen2002 with M(ensemble size) = 3, 7, 11, 21.	86
4.10	Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with M(ensemble size) = 3, 7, 11, 21.	86
4.11	Synthetic noisy data sets created by adding artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% to Chowdary2006	89
4.12	Synthetic noisy data sets created by adding artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% to Chen2002	89
4.13	Noise Robustness with artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% added to Chowdary2006 and Chen2002.	89