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

Chapter 1

Introduction

1.1 Clustering and Bioinformatics

Bioinformatics is a relatively new and fast growing field that uses computational methods to solve biology problems. Its primary goal is to increase our understanding of biological processes by developing and applying computationally intensive techniques [209]. A working definition of Bioinformatics provided by the Biomedical Information Science and Technology Initiative Consortium (BISTIC) of the US National Institutes of Health (NIH) is 'Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data' [247]. Figure 1.1 shows various disciplines involved in bioinformatics. Figure 1.2 shows applications of bioinformatics tools in various areas of biological sciences. Figure 1.3 shows an example of clustering in bioinformatics.

The rapid development of biological technologies in the past few decades lead to the exponential growth of the amount of biological data, including genomic data and gene expression data [132]. Such enormous amount of biological data raises a main challenge in bioinformatics: how to intelligently extract useful information from these data. Solving this challenge requires developing of tools and methods capable of transforming all these heterogeneous data into biological knowledge [160]. Such biological knowledge include: time and place of gene expression during development, and physiological response and disease. Traditional gene-by-gene approaches are not

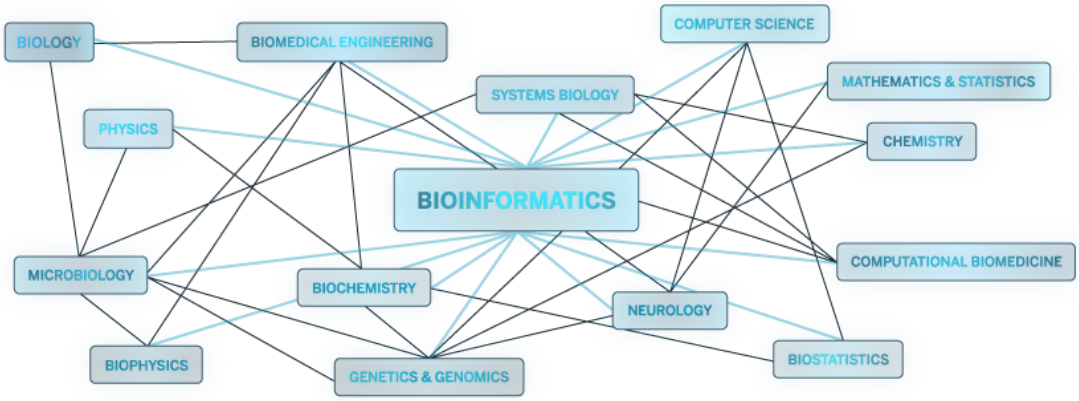

Figure 1.1: Bioinformatics and related disciplines [267]

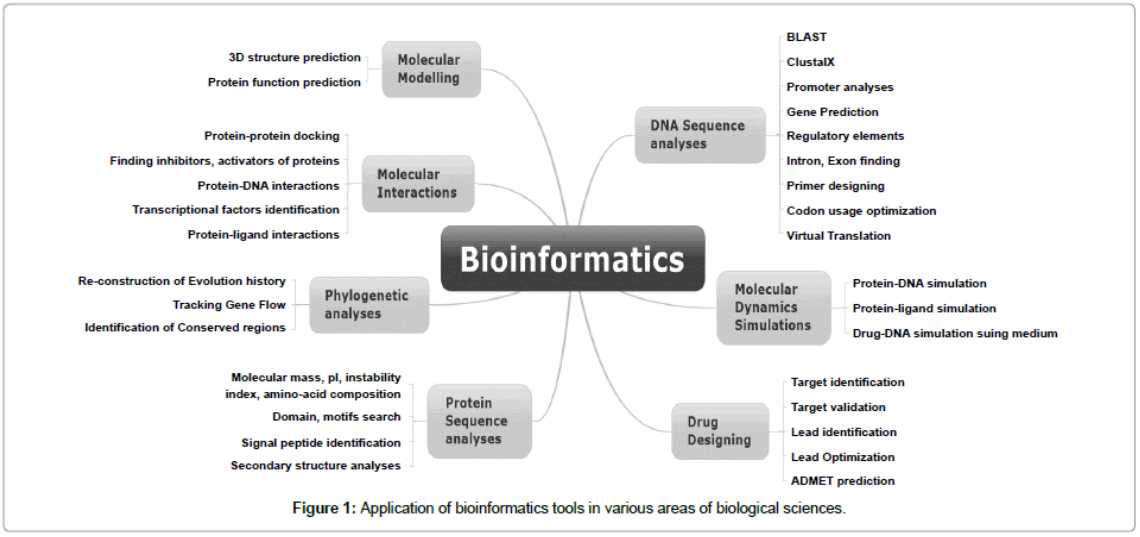

Figure 1.2: Bioinformatics applications [191]

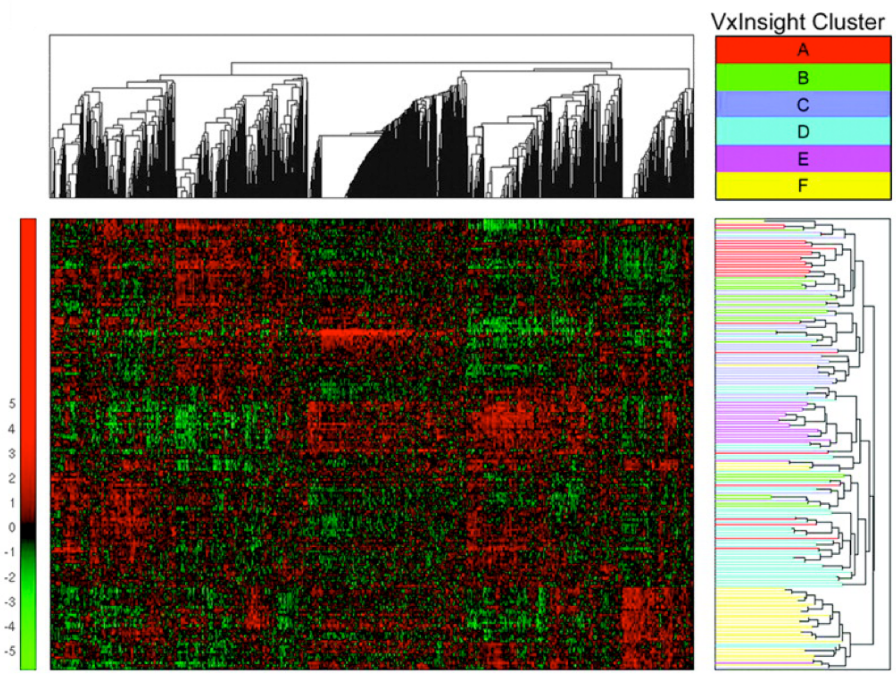

Figure 1.3: Clustering in bioinformatics [184]

sufficient [132] or practical anymore. High throughput analysis methods such as machine learning techniques are increasingly being utilized to address problems in bioinformatics, and they are producing promising results.

Gene expression is a process through which the coded information of a gene is converted into structures that operate in the cell [249]. Expressed genes include genes transcribed into mRNAs and then translated into protein. It also include genes transcribed into RNA but not translated into protein such as tRNAs and rRNAs [135, 195]. Traditional genomic research focuses on the local examination and collection of data on single genes [132]. Modern microarray technology can measure the expression levels of thousands of genes at the same time [226]. DNA microarrays usually consist of thin glass or nylon substrates containing specific DNA gene samples spotted in an array by a robotic printing device [119]. After fluorescently labeled mRNA from an experimental condition are spread onto the array, they binds or hybridizes strongly with some DNA gene samples but weakly with others depending on the inherent double helical characteristics. The array and sensors are scanned by a laser to detect

the fluorescence levels using red and green dyes. Such fluorescence levels indicate how strong each gene expresses in such experimental condition. The logarithmic ratio of the intensity of each dye is used as the gene expression value [118].

The gene expression data obtained from a scanning process contains noise. There are multiple steps to obtain microarray data, due to several system or design issues each step may introduce noise. The noise can come from five phases of data acquisition: microarray manufacturing, preparation of mRNA from biological samples, hybridization, scanning, and imaging [49]. And they can be classified into three major categories: 1) biological. cells from different populations, tissues, conditions, etc. 2) experimental. defects of the spotting equipment, different hybridization conditions and dyes, different methods to make the arrays, to culture the cells, to extract mRNA, etc. 3) processing. errors related to numerical values collection such as fluorescence scanning, image analysis, and intensity readout [299]. In general, biologic variation is the major source of variation in gene expression experiments. Noise can obscure or mislead the underlying biological meanings, which is an important reason why statistical tools are used to analyze microarray data since they can take the noise or variations into account. The noise may obscure clustering results especially in those approaches based on distance functions. Despite many pre-processing techniques can be utilized to address the noise problem, such as preforming a logarithmic transformation of each expression level or standardizing the gene expression matrix with a mean of zero and a variance of one, noise problem remains challenging in bioinformatics applications [132].

Clustering is the task of assigning a set of objects into groups or clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters [106]. Clustering is a type of unsupervised learning, which means that it does not need predetermined labels and training data set during the learning process. Clustering has a long history, tracing back to Aristotle, and has been studied extensively since the 18th century [115]. It has a wide range of applications in natural sciences, engineering, economics, marketing, medicine, psychology, and many other

fields. As a consequence, the cluster analysis literature is vast and heterogeneous with hundreds of papers and books published each year from various communities. Clustering algorithms draw upon statistics, mathematics, and computer science [115].

Many clustering algorithms use similarity measure to evaluate similarity between two data objects. Euclidean distance is one of the most commonly used similarity measure. It is the square root of the sum of the squares or the differences between the respective coordinates in each of the dimensions. However in the case of gene expression data, the overall data patterns are more relevant than individual values of each feature. So each feature vector is standardized or rescaled to have a a mean of zero and a variance of one before calculating the distance. An alternate similarity measure is Pearson's correlation coefficient. It is the covariance of the variables divided by the product of their standard deviations. It is extensively used and has been demonstrated as an effective similarity measure for gene expression data. However, previous studies have shown that it is not robust against noise. After data standardization, there is consistency between Pearson's correlation coefficient and Euclidean distance. Therefore, the effectiveness of a clustering algorithm is expected to be equivalent when either is chosen as the similarity measure [132].

With gene expression data, clustering either genes or samples is meaningful. When samples are clustered into groups, each group may correspond to some particular macroscopic phenotype, such as clinical syndromes including cancer types or subtypes.

Microarray gene expression data has been used for cancer subtyping [9, 104, 246], and the results are promising with improved accuracy over traditional methods [246]. This kind of analysis was first employed in [104] and [8]. Since then, clinical decision support in the form of cancer subtype diagnosis based on microarray data analysis has become an important emerging medical application, and has attracted great attention [69].

There are several categories regarding clustering on cancer gene expression data. The first category is: sample-based clustering. It can be used for cancer diagnosis,

cancer subtype diagnosis. The second category is: gene-based clustering. It can be used for pathway identification, gene signature discovery. The third category is: Bi-clustering. It can be used for pathway identification, functional biology.

There are several steps involved in cluster ensemble. The first step is ensemble generation. In this step, we generate a number of clustering algorithms. One option is homogeneous generation which means we use the same algorithm but different parameters, and another option is heterogeneous generation in which we use different clustering algorithms. The second step is ensemble pruning. We remove some of the clustering algorithms that may not be necessary or may not be appropriate. There are different approaches to remove such algorithms, such as exponential pruning, randomized pruning, and sequential pruning. The third step is ensemble consensus, which involve consensus functions, constant weighting or non-constant weighting.

There are many tools that can be used for clustering cancer gene expression data. For example programming frameworks, such as R and Matlab. And Integrative web-based tools, such as GEPAS, Expression Profiler, ASTERIAS, EzArray, CARMAweb, MAGMA, ArrayPipe, RACE, WebArray, MIDAW, ArrayMining etc.

1.2 Motivation

Gene expression microarray data has significant applications in biomedicine. Its enormous quantity require effective analysis approaches. A main technique of microarray data analysis is clustering such data into biological or medical meaningful groups based on their pattern of expression. However, due to the specific characteristic of gene expression data such as noisy and high dimensional, as well as the special requirements from the domain of biology, clustering gene expression data is still facing challenges. Effective clustering methods have been demanded to overcome the challenges [297].

Research on gene expression in cancer has been advancing, in particular in investigating the possibility of using gene expression data to improve the accuracy of cancer patient classification. The ability to accurately classify cancer patients into

risk groups, i.e. to predict the prognosis on an individual basis, is a key factor in making therapeutic decisions. This is especially critical for cancer therapies due to their serious side effects. Therefore, the classification of cancer patients into risk groups is a very active field of research, and it has direct clinical applications [183].

Many clustering methods have been designed and applied to cancer gene expression data for the purpose of cancer classification. They aim to improve therapeutic results by diagnosing cancer types or subtypes with improved accuracy in comparison with traditional methods such as Histopathology or Immunohistochemistry.

Cancer is a leading cause of death worldwide, which accounts for 8.2 million deaths in 2012 [276]. Annual cancer cases is expected to rise about 50% from 14 million in 2012 to 22 million in the next 20 years. Premature death and disability caused by cancer has a greater economic impact than all causes of death [252]. Despite enormous efforts in combating cancer, survival rates remain low in most forms of cancers. The problem is that conventional cancer therapy provides treatment according to less accurate cancer diagnosis methods i.e. the organ or tissue in which the cancer originates. Accurate early diagnosis (thus proper early treatment) is crucial in treating cancer. Traditional cancer diagnosis methods that are based on morphological appearance of tumors and clinical parameters do not provide enough accuracy in diagnosis.

DNA microarray technologies produce vast amount of data which are not practical or even possible to be analyzed manually. Machine Learning (ML) methods have been used to automatically analyze such microarray data and they are having a significant impact on cancer research. A common and exploratory analysis is to perform clustering on the cancer/patient samples (tissues). The aim is to find group of samples sharing similar expression patterns, which can lead to discovery of new cancer subtypes. Such kind of analysis was first carried out in [104] and [8] with promising results. Since then, clustering methods have become popular in the gene expression analysis scientific community. In addition, bioinformaticians have been proposing novel clustering methods that take intrinsic characteristics of gene

expression data into account, such as noise and high-dimensionality, to improve the clustering results [36, 170, 189]. However, different algorithms (or even the same algorithm with different parameters) often provide distinct clusterings. As a result, it is extremely difficult for users to decide which algorithm and parameters will be optimal for a given set of data set for a particular task. There is no single clustering algorithm that can perform the best for all data sets [156], and discovering all types of cluster shapes and structures presented in data is impossible for any known clustering algorithm [75, 113]. Cluster ensemble or consensus clusterings recently have emerged as simple, effective, on-stop methods for improving the robustness and quality of clustering results.

1.3 Methodology

The proposed clustering method is aimed to address some of the most challenging issues with gene expression data analysis: noise and high dimensions.

The method takes form as a new type of cluster ensemble. The cluster ensemble uses an improved fuzzy clustering algorithm with different initializations as its base clusterings. The improved fuzzy clustering algorithm employs new weighted fuzzy techniques in computing cluster centers and assigning feature vectors, which is more suitable for noisy data and high dimensional data in general.

The method uses plurality voting as its consensus function to obtain the final consensus clustering using clustering results from its base clusterings. With plurality voting, each data object votes for or is assigned to one cluster in each base clustering, and the cluster who has more votes (plurality) than any other cluster is the winner.

1.4 Contributions

Our research and this dissertation contribute to the research community in the following areas: 1) proposes an improved fuzzy clustering ensemble methodology in addressing important challenges in gene expression data analysis: noisy data and

high dimensional data; 2) provides extensive performance study on both real world cancer gene expression data sets and synthetic noisy data sets - the experimental results indicate that our approach is effective; 3) provides comprehensive review of traditional and state-of-art clustering algorithms; and 4) provides directions for future research from clustering algorithm perspective and bioinformatics perspective.

1.5 Dissertation Organization

The remainder of this dissertation is structured as follows: Chapter 2 gives relevant background information and literature review on clustering methods and their applications on gene expression data. Chapter 3 presents the proposed methodology: Improved Fuzzy Cluster Ensemble (IFCE). Chapter 4 details experiments and evaluates the proposed methodology. Conclusions and future research directions are provided in Chapter 5.

Chapter 2

Background and Literature Review

2.1 Introduction

Clustering bio-molecular data has been used to improve cancer subtyping [9, 104] over traditional clinical methods based on morphological appearances. Its aim is to find groups of patients sharing similar expression patterns or biological attributes. Due to the amount of data, e.g. large number of dimensions or gene expression data produced by microarray technology, manual analysis is not possible. Automatic analyzing tools are needed to discover underlying patterns within the data. Clustering approaches are suitable to accomplish this goal and have shown promising progress and possibilities for more accurate and reliable results.

In order to develop clinically successful clustering-based cancer subtyping tools for microarray data, a solid understanding of clustering and the available clustering methods is essential. Clustering has a long history, tracing back to Aristotle, and has been studied extensively since the 18th century [115]. Clustering is the task of assigning a set of objects into groups or clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters. It has a wide range of applications in natural sciences, engineering, economics, marketing, medicine, psychology, and many other fields. As a consequence, the cluster analysis literature is vast and heterogeneous with hundreds of papers and books published each year from various communities.

Cluster analysis algorithms draw upon statistics, mathematics, and computer

science [115]. Closely related fields are machine learning, pattern recognition, computer vision, image analysis, information retrieval, and bioinformatics. The k-means algorithm, first proposed in 1957 [171] (it wasn't published outside Bell Labs until 1982), is one of the most simple and popular clustering algorithms. Thousands of clustering algorithms in various fields have been published since then. Due to the ill-posed nature of clustering, *i.e.* lack of external objective criteria to validate clustering results, it is difficult to design a general purpose clustering algorithm. Different clustering algorithms or even the same algorithm with different parameters often produce different results on the same data set. There is no single clustering algorithm that performs best for all data sets, and discovering all cluster structures in a data set is impossible for any known clustering algorithm [75, 113].

In order for computational communities to contribute to the cancer subtyping field, the background and an updated knowledge of cancer subtyping are necessary. Cancer remains a leading cause of death worldwide largely due to lack of effective treatment. Personalized treatment based on cancer subtypes improves patient survival. The goal of cancer subtyping is to identify subtypes within a cancer type, where patients within a subtype are more similar than patients in other subtypes. The advent of microarray technology in the 1990s made it possible to assess the expression of tens of thousands of genes in a single experiment. Microarray gene expression data has been used for cancer subtyping [9, 104, 246], and the results are promising with improved accuracy over traditional methods [246, 104, 8]. Since then, clinical decision support in the form of cancer subtype diagnosis based on microarray data analysis has become an important emerging medical application, and has attracted great attention [69].

The high dimensional and noisy nature of gene expression data has given rise to a wealth of clustering techniques being presented. Much of the early work used methods developed originally for other domains [8, 28, 30, 44, 78, 104, 159, 167, 244, 294]. Novel algorithms specifically targeting gene expression data and taking its intrinsic characteristics into account were presented to improve the clustering results [208, 36,

197].

The main goal of this review is to provide a background of cluster analysis application in cancer subtyping, as well as an overview of its current state. It is important to understand the difference between clustering (unsupervised learning) and classification (supervised learning). In contrast to classification techniques, clustering techniques do not require labels which may not be accurate or available. Clustering results are obtained solely from data. This advantage also enables clustering algorithms to avoid over-fitting, a potential problem in classification techniques.

Audiences in the medical community may find that the overview of clustering and literature review in the computational communities will broaden their experimental possibilities, while the audience in the computational communities may find the cancer subtyping background and literature review in the clinical community an entry point for them to start contributing to this application area. Since clustering is a vast and ever changing field, it is impossible to cover all approaches in a single paper. This review paper focuses on key clustering algorithms and their novelties. It also covers important developments applying clustering techniques to cancer subtyping.

The organization of the paper is as follows: Section 2.2 presents a general overview of clustering, including discussions about issues such as the curse of dimensionality, feature selection, and cluster validity. It then reviews literature regarding different types of clustering across several disciplines including their variety, uses, strengths, and limitations. Section 2.3 provides a literature review of clustering applications on microarray-data-based cancer subtyping. It includes literature in both the clinical community and the computational community. For each community, the section discusses literature in two categories: mRNA experiments and miRNA experiments. Section 2.4 examines challenges in microarray-data-based cancer subtyping. Finally, Section 2.5 concludes the review and points out future directions.

2.2 Data Clustering

Clustering is an interdisciplinary research topic and is also known by researchers in different fields as unsupervised learning, exploratory data analysis, grouping, clumping, taxonomy, typology, and Q-analysis [129]. Some of its applications include numerical taxonomy [125], class discovery [285], and natural classification [129]. Clustering has become increasingly popular as the society increasingly generates an overwhelming amount of data, and it is often used as the first step in data analysis or as a preparation step for experimental work.

There is no universal agreement upon definition of clusters. A cluster is a set of objects that are compact (or similar to each other) and isolated (or dissimilar) from other clusters. In reality, cluster definition is subjective, and its significance and interpretation requires related domain knowledge [129]. Similarity measure is used by clustering methods to calculate the similarity between two objects. Different similarity measures will have different clustering results, as some objects may be similar to one another using one measure but dissimilar using another. Similarity between two objects can be measured in different ways, and the three dominant methods are distance measures, correlation measures, and association measures. Common similarity measures include Euclidean distance, Manhattan distance, Maximum norm, Mahalanobis distance, Pearson coefficient, Spearman's rank correlation coefficient, angle between two vectors, and the Hamming distance.

There is no single clustering algorithm that performs best across all problems or data sets. Therefore, it is important to study the characteristics of the problem and use an appropriate clustering strategy [285]. Properties to be considered in choosing a clustering algorithm include [26]: a) feature type (numeric and non-numeric), b) scalability (large datasets), c) handling high dimensional data, d) finding clusters of irregular shape, e) handling outliers, f) time complexity of the algorithm, g) data order dependency, h) assignment type (hard or strict vs. soft or fuzzy), i) prior knowledge and user defined parameters dependency, and j) interpretability and visulization of

results.

Clustering techniques can be organized into categories. Different criteria may result in different categories of clustering algorithms [285]. Furthermore, categorization of clustering algorithms is not straightforward or canonical, and categories can overlap [26]. For convenience, in this review we use the following taxonomy, which is also widely used in the literature: hierarchical clustering (Section 2.2.1), partitioning clustering (Section 2.2.2), graph-based Clustering (Section 2.2.3), distribution-based clustering (Section 2.2.4), density-based clustering (Section 2.2.5), grid-based clustering (Section 2.2.6), clustering high dimensional data (Section 2.2.7), and other clustering techniques (Section 2.2.8).

2.2.1 Hierarchical Clustering

Hierarchical clustering algorithms organize a data set into a hierarchical structure according to a similarity measure. These algorithms connect objects based on their similarity to form clusters, which is usually represented using a dendrogram. Hierarchical clustering algorithms differ in the choice of similarity measures, the linkage criterion (distance between clusters), and whether the process is agglomerative (bottom-up) or divisive (top-down). Agglomerative hierarchical clustering starts with singleton clusters and then recursively merges appropriate clusters, and divisive hierarchical clustering starts with one cluster containing all objects and recursively splits appropriate clusters [26, 285].

Two classical divisive approaches are DIANA [142]. DIANA (DIvisive ANAlysis Clustering) selects in each dividing step the cluster with the largest diameter and divides it into two new clusters. DIANA chooses the object from this largest cluster with the maximum average dissimilarity and then moves all objects to the cluster that are more similar to the new cluster than to the remainder. MONA (MONothetic Analysis Clustering of Binary Variables) can only be applied when the variables are binary, in which case splits can be made using one variable at a time. MONO takes individual variables in sequence rather than amalgamating them into an average. Other heuris-

tic divisive approaches are conceptually more complex than agglomerative clustering since we needs a second clustering algorithm, such as k-means, to divide the data at each step. Most divisive approaches can be generalized as a recursive application of partitioning-based clustering (see Section 2.2.2).

Agglomerative clustering algorithms typically define a linkage (distance between two clusters) and then selects in each merging step the two clusters that have the shortest distance. There are many classical agglomerative hierarchical clustering algorithms based on different linkage criterion [251, 289, 200]. The single linkage method or nearest neighbor methods [105, 243] use the distance between two closest objects in different clusters. The complete linkage methods [66] use the distance between two farthest objects in different clusters. The average linkage methods uses average (or weighted average) distance between two objects in different clusters [253]. The centroid or median linkage methods [251] uses Euclidean distance between unweighted centroids of different clusters. Ward's method [275] considers the relationship of all objects in a cluster. Its objective is to form clusters such that the increase of variance within each group is minimized.

Figure 2.1 shows an example of hierarchical clustering using single (top panel) and average linkage (bottom panel) on the same gene expression data. The data is a numerical matrix where the rows represent different samples (patients) and the columns represent different genes. A hierarchical clustering can be applied on both gene and sample space. The structure of the tree, as well as the clustering results, is highly dependent on the linkage criterion. For example, cutting the tree of the samples (rows) built for average linkage into two will result in two clusters with approximately equal sizes. However, cutting the tree build for single linkage will result in a large cluster (with most samples) and one small cluster with only one sample.

Advantages of hierarchical clustering include: a) good visualization with dendrogram representation, b) very informative descriptions with dendrogram representation, and c) flexibility regarding the number of clusters (the clustering results can be obtained by cutting the dendrogram at different levels). Disadvantages of hierarchi-

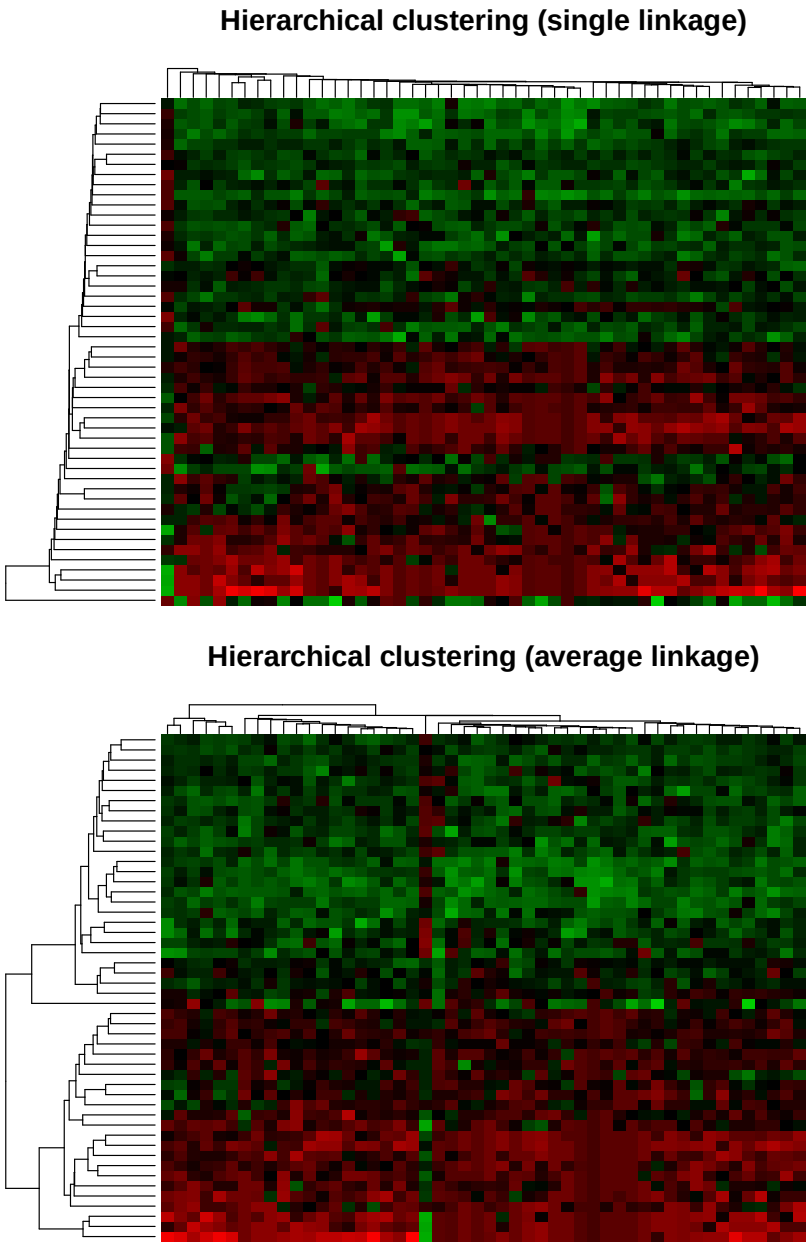

Figure 2.1: An example of hierarchical clustering applied on gene expression data. In this example, the rows represent the genes while the columns represent different samples. The expression values are color coded (from red to green). The hierarchical clustering are performed on both rows (genes) and columns (samples).

cal clustering include: a) lacking of robustness and sensitivity to noise and outliers (small changes in data or linkage might result in completely different results), b) high computational complexity, which is typically $\mathrm{O}(n^3)$ for agglomerative algorithms and $\mathrm{O}(n^2)$ for divisive algorithms, and c) prone to reversal phenomenon, i.e. two clusters being merged at some step are closer to each other than pairs of clusters merged earlier.

2.2.2 Partitioning Clustering

Partitioning clustering algorithms divide objects into clusters without hierarchical structure. Clusters are represented by a central vector. Given the number of clusters, partitioning clustering assigns the objects to the closest cluster center. Partitioning algorithms can be grouped into k-means methods and k-medoids methods. k-means methods use the centroid of objects within a cluster as center. k-medoids methods use the most appropriate object within a cluster as center.

The k-means clustering [26, 88, 116, 178] is one of the most widely used clustering algorithms. There are many variations of the basic k-means clustering. Classic k-means reassigns data objects based on minimizing the residual sum of square (RSS): $RSS = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$ where k is the number of clusters and μ_i is the mean of points in cluster S_i. Given a predefined number of clusters, the algorithm starts by choosing k initial centers and assign each object to the nearest center. At each iteration, the centers are recomputed and reassign each object to the new centers. The iterations stop when no further update is needed. FORGY [88] reassigns objects to nearest centroids and recomputes centroids. ISODATA [20] (Iterative Self-Organizing Data Analysis Technique) splits and merges intermediate clusters based on a user-defined threshold and iterates until the threshold is reached. It iterates until a stopping criterion is achieved. Fuzzy c-means [27, 77] assigns fuzzy cluster membership to each data object, and updates cluster centers and membership after each iteration. Methods to speed up k-means and fuzzy C-means such as brFCM (bit reduction by Fuzzy C-Means) [81] replace similar data objects with their centroid

before clustering.

Variations of k-medoid [142] methods are as follows. PAM (Partitioning Around Medoids) assigns each data object to the closest medoid and iteratively reassigns objects and updates medoids to optimize the objective function. CLARA (Clustering LARge Applications) [142] applies PAM on multiple subsets or samples of the data set, and selects the best clustering as output. CLARANS (Clustering Large Applications based upon RANdomized Search) [203] searches a graph where each node is a set of medoids. It selects a node randomly in search for a local minimum among its neighbor nodes through iterations and outputs the best node to form clustering results.

Advantages of partitioning clustering include: a) simple, straightforward and easy implementation, b) fast execution with computation complexity of $\mathrm{O}(n)$, c) very suitable for compact and hyperspherical clusters, d) computational rigor (firm foundation of analysis of variances). Disadvantages of partitioning clustering include: a) they are still subjective processes that are sensitive to assumptions, b) they require the number of clusters to be specified in advance, c) they prefer clusters of approximately similar size, as they will always assign an object to the nearest center, often leading to incorrectly cut borders in between of clusters, d) they are subject to easy trapping in local minima and sensitivity to the initial partition (hill-climbing optimization method).

Figure 2.2 shows an example of partitioning clustering applied on a lung cancer dataset GSE19188 [122] downloaded from Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo`). The data is visualized in the fist 3 principal components. The algorithm, named PINS (Perturbation clustering for data INtegration and disease Subtyping), is built on the basic k-means algorithm. PINS repeatedly perturbs the data (slightly change the expression values) and cluster the patients using different values of k (number of clusters) using Hartigan and Wong's algorithm [116]. It then choose the partitioning that is the most robust to data perturbation.

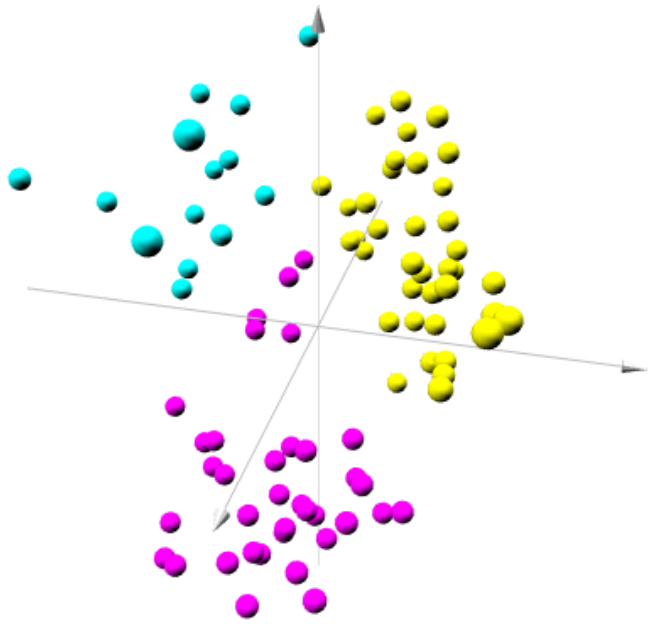

Figure 2.2: An example of k-means based clustering [208] on a lung cancer dataset [122]. The data shown in the space of the first three principal components. Different colors represent different clusters.

2.2.3 Graph-based Clustering

Graph-based clustering algorithms construct a graph/hypergraph from the data and then partition the graph/hypergraph into subgraphs/subhypergraphs or clusters. Each vertex represents a data object, and the edge weight represents the similarity of two vertices [43]. The edges in the same subgraph/subhypergraph should have high weights, and the edges between different subgraphs/subhypergraph should have low weights [43]. It is also called spectral clustering [129].

Classical graph-based algorithms are as follows. Chameleon [140] uses a connectivity graph and graph partitioning to build small clusters, followed by the agglomerative hierarchical clustering process. Its key feature is that it considers both interconnectivity and closeness when merging clusters. CACTUS (Clustering Categorical Data Using Summaries) [93] detects candidate clusters based on the summary of the data set and determines the actual clusters through a validation process against the candidate clusters. It uses a similarity graph to represent the inter-attribute and

intra-attribute summaries. A Dynamic System-based Approach or STIRR (Sieving Through Iterated Relational Reinforcement) [100] represents each attribute value as a weighted vertex in a graph. It iteratively assigns and propagates weights until a fixed point is reached. Different weight groups correspond to different clusters on the attribute. ROCK (Robust Clustering algorithm for Categorical Data) [107] repeatedly merges two clusters until the specified number of clusters is reached, and it uses data sampling to improve complexity. It uses a connectivity graph to calculate the similarities between data objects.

The advantages of graph-based clustering include [43]: a) a graph is an elegant data structure that can model many real applications, b) it is based on solid mathematical foundations, including spectral theory and Markov stochastic process, and c) it produces optimal clustering (optimizing a quality measure instead of acting greedily toward the final clustering). The major disadvantage of graph-based clustering is that it may be slow when working on large scale graphs [43]. In addition, the partitioning results highly depend on the way the graph is constructed from the raw data [208].

2.2.4 Distribution-based Clustering

Distribution-based clustering views or assumes that the data are generated by a mixture of probability distributions, each of which represents a different cluster [92, 188]. This way, a cluster can be seen as objects generated by the same distribution. Thus, a particular clustering method can be expected to produce good results when the data conform to the method's distribution model [92]. It is also called model-based clustering. There are usually two approaches to form the model: the classification likelihood approach and the mixture likelihood approach [92].

Classical distribution-based algorithms are as follows. The EM (Expectation-Maximization) clustering algorithm [68] is the most popular method in distribution-based clustering. It tries to fit the data set into the assumed number of Gaussian distributions by moving the means of Gaussian distributions toward the cluster centers.

COOLCAT (reducing the entropy, or COOLing of the CATegorical data clusters)[23] uses entropy to cluster categorical data. It consists of data sampling and incremental assignment. STUCCO (Search and Testing for Understandable Consistent Contrasts) [24] uses tree searching and significant contrast-sets to find clusters. GMDD (Gaussian Mixture Density Decomposition) [301] uses a recursive approach and identifies each Gaussian component in the mixture successively. Autoclass [42] is based on the classic distribution-based approach and uses a Bayesian method to determine the optimal clusters. P-AutoClass [221] is a parallel version of Autoclass and can be used on large data sets.

The advantages of distribution-based clustering include [26]: a) it can be modified to handle complex data, b) it has a solid theoretical foundation, c) Its results are easily interpretable, d) it not only provides clusters, but also produce complex models that capture relationships among attributes, e) results are independent of the timing of consecutive batches of data, f) it is good for online learning since the intermediate mixture model can be used to cluster objects, and g) the Mixture model can be naturally generalized to cluster heterogeneous data. The disadvantage of distribution-based clustering is the difficulty in choosing the appropriate model complexity (since a more complex model will usually be able to explain the data better but may cause an overfitting problem from excessive parameter set).

2.2.5 Density-based Clustering

Density-based clustering defines clusters as dense regions of data objects separated by low-density regions. A cluster is a connected dense component and grows in any direction that density leads [92]. Objects in low-density areas which separate clusters are usually considered to be noise and border points. There are two major approaches for density-based clustering [26]: the connectivity approach pins density to a training data point; the density function approach pins density to a point in the attribute space.

Representative algorithms for the connectivity approach are as follows. DBSCAN

(Density-Based Spatial Clustering of Applications with Noise) [82] starts by selecting a data object and tries to find all data objects density-reachable from it to form a cluster. If none are found, the algorithm selects a new data point and repeats. GDBSCAN (Generalized DBSCAN) [233] generalizes the concept of neighborhood by permitting the use of any distance function besides Euclidian distance and allows other measures besides simply counting the objects to define the cardinality of that neighborhood. OPTICS (Ordering Points To Identify the Clustering Structure) [13] is like an extended DBSCAN algorithm. It does not assign cluster memberships but stores the order in which the data objects are processed as well as the core-distance and a reachability-distance for each data object. An extended DBSCAN is used to assign cluster memberships. DBCLASD (Distribution Based Clustering of LArge Spatial Databases) [288] uses the notion of clusters based on the distance distribution and incrementally augments an initial cluster by its neighboring points as long as the nearest neighbor distance set of the resulting cluster still fits the expected distance distribution.

The advantages of density-based clustering are as follows: a) they can find clusters of arbitrary shapes, in contrast to many other methods, b) their time complexity is low (linear or $\mathrm{O}(n)$), c) it is deterministic for core and noise points (but not for border points) and thus there is no need to run it multiple times. d) they can handle noise well [92], e) the number of clusters is not required since it finds clusters and the number of clusters automatically [92], f) results are independent of data ordering [26], and g) there are no limitations on the dimension or attribute types [26].

The disadvantages of density-based clustering are as follows: a) It is often difficult to detect cluster boarders when the cluster density decreases continuously (*i.e.* arbitrary borders). b) For a mixtures of Gaussians data set, distribution-based clustering (*e.g.* EM) usually outperforms density-based clustering. c) Limitations in processing high-dimensional data, since it is difficult to distinguish high-density regions from low-density regions when the data is high-dimensional [129]. d) Most density-based clustering algorithms were developed for spatial data [92].

2.2.6 Grid-based Clustering

Grid-based clustering operates on space partitioning instead of data partitioning to produce clusters [26]. It first creates the grid structure by partitioning the data space into cells (or cubes) and then clusters the cells based on their densities.

Representative algorithms are as follows. BANG-clustering [26, 235] uses a multi-dimensional grid data structure to organize or partition the data. It uses the cell information in the grid and clusters the cells. STING (A STatistical INformation Grid approach) [273] uses a hierarchical structure of grid cells with a top-down approach. It labels a cell to be relevant or not at a specified confidence level. Then, it finds all the regions formed by relevant cells. STING+ [26, 274] uses a similar hierarchical cell structure as STING and introduces an active spatial data mining approach. OptiGrid (Optimal Grid) [120] constructs an optimal grid partitioning of the data by finding the best partitioning hyperplanes for each dimension with projections of the data. GRIDCLUS (GRID-CLUStering) [234] organizes the space surrounding the clusters with a grid data structure. It uses a topological neighbor search to cluster the grid cells. GDILC (Grid-based Density-IsoLine Clustering) [291] is based on the idea that the density-isoline figure reflects the distribution of data. It uses a grid-based approach to calculate the density and finds dense regions. WaveCluster (Wavelet-based clustering) [239] transforms the original feature space by applying wavelet transform and then finds the dense regions in the new space. It yields sets of clusters at different resolutions and scales, which can be chosen based on the user's needs. FC (Fractal Clustering) [22] adds one data object at a time to one cluster in such a way that the fractal dimension changes the least after adding the data object.

The advantages of grid-based clustering are as follows: a) it is fast and works well with large data sets (since speed is independent of the number of objects in the data) [26, 92]. b) it handles noise well [26]. c) it is independent of data ordering [26]. d) it can handle attributes of different types [26]. e) it can be used as an intermediate step in many other algorithms such as CLIQUE and MAFIA [26]. The disadvantages

of grid-based clustering are as follows: a) most algorithms need the user to specify grid size or density thresholds, which can be difficult (fine grid sizes result in high computational time, while coarse grid sizes result in low quality of clusters) [92]. b) some grid-based clustering algorithms (*e.g.* STING, WaveCluster) are not good at high dimensional data [92].

2.2.7 Clustering High Dimensional Data

High Dimensional Data clustering refers to clustering on data objects that represent from a few dozen to thousands or more features. Such high dimensional data are often seen in areas such as medicine (*e.g.* microarray experiments), and text documents (*e.g.* word-frequency vector methods [40]). Clustering high dimensional data is tremendously difficult. One problem is that increased irrelevant features eliminate the likelihood of clustering tendency [26]. Another problem is the 'curse of dimensionality', or lack of data separation, in high dimensional space (the problem becomes severe for dimensions greater than 15) [26]. Performing feature selection before applying clustering can improve the first problem. Principal Component Analysis (PCA) [215] is commonly used. However, the dimension may still be high after feature selection. In this review, we discuss techniques that have been developed to address such situations: projected clustering, subspace clustering, bi-clustering (or co-clustering), tri-clustering, hybrid approaches, and correlation clustering.

Projected Clustering:

Projection techniques map data objects from a high dimensional space to a low dimensional space, while maintaining some of the original data's characteristics [15].

Examples are as follows. PreDeCon [29] finds subsets of feature vectors that have low variance along subsets of attributes. PROCLUS [4] finds the candidate clusters and dimensions by using medoids. For each medoid, the subspace is determined based on attributes with low variance. Random projections for k-means clustering [35] implements a dimensionality reduction technique for k-means clustering based on

random projections.

Subspace Clustering:

Subspace clustering algorithms identify clusters in appropriate subspaces of the original data space.

Examples are as follows. CLIQUE (CLustering In QUEst) [6] partitions the data space into units and then finds the maximum sets of connected dense units. SUBCLU (density-connected Subspace Clustering) [152] adopts the notion of density-connectivity introduced in DBSCAN (Section 2.2.5) and uses the monotonicity of density-connectivity to prune subspaces. CACTUS (Clustering Categorical Data Using Summaries) is covered in Section 2.2.3. ENCLUS (ENtropy-based CLUStering) [46] finds clusters in subspaces based on entropy values of subspaces. Subspaces with lower entropy values typically have clusters. It then applies CLIQUE or other clustering algorithms to such subspaces. MAFIA (Merging of Adaptive Finite Intervals) [103] uses adaptive grids in each dimension and then merges them to find clusters in higher dimensions. OptiGrid (Optimal Grid) is covered in Section 2.2.6. MrCC (Multi-resolution Correlation Cluster detection) [55] constructs a novel data structure based on multi-resolution and detects correlation clusters by identifying initial clusters as axis-parallel hyper-rectangles with high data densities, followed by merging overlapping initial clusters.

Hybrid Approaches:

Hybrid approaches find overlapping clusters. Some of them find only potentially interesting subspaces and use full-dimensional clustering algorithms to obtain the final clusters.

Examples are as follows. DOC (Density-based Optimal projective Clustering) [225] uses a global density threshold to compute an approximation of an optimal projective cluster. FIRES (FIlter REfinement Subspace clustering) [153] first computes one-dimensional clusters and then merges them by applying 'clustering of clusters'

based on the number of intersecting points between clusters. P3C (Projected Clustering via Cluster Cores) [196] first computes intervals matching or approximating higher-dimensional subspace clusters on every dimension and then aggregates those intervals into cluster cores. The cluster cores are refined and used to assign data objects.

Bi-clustering:

Bi-clustering is also called bi-dimensional clustering [47], co-clustering, coupled clustering, or bimodal clustering. Bi-clustering is popular in bioinformatics research, especially in gene or sample clustering. For gene expression data, there are experimental conditions in which the activity of genes is uncorrelated. This causes limitations for results obtained by standard clustering methods. So bi-clustering algorithms that can perform simultaneous clustering on the genes and conditions are developed to find subgroups of genes and subgroups of conditions in which the genes exhibit highly correlated activities for every condition [179].

Examples are as follows. CTWC (Coupled Two-Way Clustering) [98] generates submatrices by an iterative process and considers only those submatrices whose rows and columns belong to genes and samples/conditions that were in a stable cluster in a previous iteration. ITWC (Interrelated Two-Way Clustering) [263] clusters the rows and then clusters the columns, based on each row cluster. It keeps the cluster pairs that are most dissimilar. Block Clustering [117] sorts the data by row mean or column mean and splits the rows or columns such that the variance within each 'block' is reduced. It then repeats and splits rows or columns differently. δ-biclusters [47] or CC algorithm (Cheng and Church's) finds biclusters whose rows and conditions show coherent values, using mean-squared residue. SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) [261] uses probabilistic modeling and graph theoretic techniques to find subsets of rows whose values are very different in a subset of columns. Plaid Models [161] allows biclusters to overlap, *i.e.* a gene or a sample/condition can belong to more than one cluster. Information-theoretic co-clustering [70] intertwines

the row and column clusterings to increase mutual information.

Correlation Clustering:

Correlation clustering uses the correlations among attributes to guide the clustering process. These correlations may be different and exist in different clusters and cannot be reduced to uncorrelated ones by traditional global decorrelation techniques. Such correlations create clusters with different spatial shapes, and local correlation patterns are used to define the similarity between data objects. Correlation clustering is closely related to biclustering.

Examples are as follows. ORCLUS (ORiented projected CLUSter generation) [5] is similar to k-means but uses a distance function based on an eigensystem, *i.e.* the distance in the projected subspace. The eigensystem is adapted during iterations and close pairs of clusters are merged. 4C (Computing Correlation Connected Clusters) [29] takes a density-based approach and uses a density criterion to grow clusters. The density criterion is the minimal number of data objects within the neighborhood of a data object. The neighborhood is based on distance between two data objects in the eigensystems. HiCO (Hierarchical COrrelation clustering) [2] defines the similarity between two data objects based on their local correlation dimensionality and subspace orientation. It takes a hierarchical density-based approach to obtain correlation clusters. CASH (Clustering in Arbitrary Subspaces based on the Hough transform) [1] is based on the Hough transform [123], which maps the data space into parameter space. It then uses a grid-based approach to find dense regions in the parameter space and corresponding data subsets in the original data space. It recursively applies itself on such corresponding data subsets.

2.2.8 Other Clustering Techniques

Neural Network-Based Clustering:

The neural network approach has been studied intensively by mathematicians, statisticians, physicists, engineers, and computer scientists [155]. A neural network is

an interconnected group of artificial neurons and an adaptive system for information processing. Neural-network-based clustering is competitive-learning-based clustering, not statistical model-identification based clustering. For competitive-learning-based clustering, the first phase is learning where the algorithmic parameters are adjusted, and the second phase is generalization [73]. Competitive learning can be implemented using a two-layer neural network: the input layer and the output layer [73].

Examples are as follows. A SOM (Self-Organizing Map) [150] consists of nodes or neurons, each of which is associated with a weight vector and a position in the map space. It creates a mapping from a higher dimensional input space to a lower dimensional output space. SOM clustering computes the distance of the input pattern to each neuron and finds the winning neuron. LVQ (Learning Vector Quantization) or VQ (Vector Quantization) [37, 96] is a classical quantization technique for signal processing. It models the probability density functions by using the distribution of prototype vectors. It divides a set of vectors into groups that have approximately the same number of vectors closest to them. Basic VQ is k-means clustering, and LVQ is a precursor to self-organizing maps (SOM) [96]. Neural gas [185] is inspired by SOM. It is a simple algorithm and finds optimal data representations based on feature vectors. During the adaptation process, the feature vectors distribute themselves dynamically like a gas within the data space.

Evolutionary Clustering:

Evolutionary computation has many applications in computer science, bioinformatics, pharmacometrics, engineering, physics, and economics. Evolutionary computation is inspired by the biological mechanisms of evolution, and uses iterative processes such as growth or development followed by selection in a population of candidate solutions. Clustering methods that use local search techniques including hill-climbing approach-based k-means suffer from local minima problems. The recent advancements in evolutionary computational technologies [87] provide an alternate and effective way to find the global or approximately global optimum [287]. PSO (Particle

Swarm Optimization) simulates social behavior in nature, such as bird flocking or fish schooling [146]. ACO (Ant Colony Optimization) algorithms model the behaviors of ants in nature [72]. GAs (Genetic Algorithms) [121] mimic natural selection and use evolutional mechanisms such as crossover, mutation and selection to generate solutions.

Examples are as follows. PPO (Particle-Pair Optimizer) [74] is a modification of the Particle Swarm Optimizer. It uses two particle pairs to search for the global optima in parallel and uses k-means for efficient clustering. Niching genetic k-means [241] modifies Deterministic Crowding [180], one of the niching genetic algorithms, and incorporates one step of k-means into its regeneration steps [241]. Evo-Cluster algorithm [177] encodes cluster structure in a chromosome, in which one gene represents one cluster or the objects belonging to one cluster. Reproduction operators are used between chromosomes. GenClust [97] is a simple algorithm and proceeds in stages. It uses genetic operators and a fitness function to compute partitions in a new stage based on partitions in the previous stage.

Kernel Clustering:

Kernel-based learning such as Support Vector Machines (SVMs) [57, 236] has had successful applications in pattern recognition and machine learning and is becoming increasingly important. Kernel methods [59] perform a non-linear mapping of the low dimensional input data into a high dimensional space, which becomes linearly separable. To improve efficiency, they avoid explicitly defining the nonlinear mapping by using kernel functions, such as polynomial kernels, sigmoid kernels, and Gaussian radial basis function (RBF) kernels. This is the known as the *kernel trick*.

Examples are as follows. SVC (Support Vector Clustering) [272, 296] uses SVM training to find the cluster boundaries and an adjacency matrix to assign a cluster label to each data object [287]. Variations of SVC include Iterative One-Class SVC [39], and rough Set SVC [214]. Kernel k-means [101] uses a kernel method to calculate the distance between items in a data set, instead of using the Euclidean

distance as in regular k-means. Variations include Incremental Kernel-k-means [237]. Kernel deterministic annealing clustering [293] uses an adaptively selected Gaussian parameter and a Gaussian kernel to determine the nonlinear mapping. Kernel fuzzy clustering [169, 298, 300] applies kernel techniques to fuzzy clustering algorithms by replacing the original Euclidean distance with a kernel-induced distance. Kernel Self-Organizing Maps [11, 34] perform self-organizing between an input data object and the corresponding prototype in the mapped high dimensional feature space or in the mapped space completely.

Sequential Data Clustering:

Sequential data are sequences of numerical data or non-numerical symbols and can be generated from speech processing, video analysis, text mining, gene sequencing, and medical diagnosis. Time series data or temporal data are a type of sequential data, which, unlike static data, contain feature values that change over time. Since sequential data usually have variable length, dynamic behaviors, and time constraints [110], they cannot be represented as points in the multi-dimensional feature space and thus cannot be analyzed using any of the clustering techniques we have mentioned thus far [287]. Clustering techniques targeting sequential data have been developed, and they commonly use three strategies: proximity-based approaches, feature-based approaches, and model-based approaches.

Proximity-based approaches use proximity information such as the distance or similarity between pairs of sequences. They then use hierarchical or partitional clustering algorithms to group the sequences into clusters [287]. Examples are as follows. The Needleman-Wunsch algorithm [201] uses basic dynamic programming and is a global optimal alignment algorithm. The Smith-Waterman algorithm [248] is based on Needleman-Wunsch algorithm, and also uses dynamic programming. It compares multi-lengthed sequence segments using character-to-character pair-wise comparisons. FASTA (FAST-All) [216] first finds segments of the two sequences that have some degree of similarity and marks these potential matches. It then performs a

more time-consuming optimized search approach such as the Smith-Waterman algorithm. BLAST (Basic Local Alignment Search Tool) [10] searches for short alignment matches between two sequences using a heuristic approach, which approximates the Smith-Waterman algorithm. GeneRage [80] automatically clusters sequence datasets by using Smith-Waterman dynamic programming alignment and single-linkage clustering. SEQOPTICS (SEQuence clustering with OPTICS) [45] implements Smith-Waterman algorithms as the distance measurement and uses OPTICS [13] to perform sequence clustering.

Feature-based approaches map sequences onto multi-dimensional data points using feature extraction methods and then use vector-based clustering algorithms on the data points [287]. Examples are as follows. Scalable sequential data clustering [109] uses a k-means based clustering algorithm which has near-linear time complexity to improve the scalability problem. Pattern-oriented hierarchical clustering [198] uses a hierarchical algorithm, which can generate the clusters as well as the clustering models based on sequential patterns found in the database. The wavelet-based anytime algorithm [271] combines a novel k-means based clustering algorithm and the multi-resolution property of wavelets. It repeatedly uses coarse clustering to obtain a clustering at a slightly finer level of approximation.

Model-based approaches assume sequences that belong to one cluster are generated from one probabilistic model [287]. Examples are as follows. Autoregressive moving average (ARMA) models [18, 283] derive an EM algorithm to learn the mixing coefficients and the parameters of the component ARMA models. They use the Bayesian information criterion (BIC) to determine the number of clusters. The Markov chain approach [228] models dynamics as Markov chains and then applies an agglomerative clustering procedure to discover a set of clusters that best capture different dynamics. The Polynomial models approach [91] assumes the underlying model is a mixture of polynomial functions. It uses an EM algorithm to estimate the cluster membership probabilities, using weighted least squares to fit the models. The Hidden Markov Model (HMM) [250] is a probabilistic model-based approach. It uses

HMMs, which have shown capabilities in modeling the structure of the generative processes underlying real-world time series data.

Ensemble Clustering:

Clustering ensembles have emerged to improve robustness, stability and accuracy of clustering results [99, 205, 207]. A cluster ensemble combines the results of multiple clustering algorithms to obtain a consensus result [220]. It can produce better average performance and avoid worst case results. Other usages of clustering ensembles include improving scalability by performing clustering on subsets of data in parallel and then combining the results, and data integration when data is distributed across multiple sources [128].

There are two main steps in a clustering ensemble: generation and consensus. In the generation step, several approaches are used [270]: different clustering algorithms, a single algorithm with different parameter initializations, different object representations, different object projections, and different subsets of objects.

In the consensus step, several approaches are used: relabeling and voting, Mutual Information (MI), co-association based functions, finite mixture models, a graph/hypergraph partitioning approach, and others.

The relabeling and voting approach is also called the direct approach. It finds the correspondence of the cluster labels among different clustering results and then uses a voting method to determine the final cluster label for a data object. Examples are as follows. BagClust1 [76] applies a clustering procedure to each bootstrap sample and obtains the final partition by plurality voting so that the majority cluster label for each data object determines the final cluster membership. BagClust2 [76] introduces a new dissimilarity matrix which contains the proportion of time each pair of data objects were clustered together in the bootstrap clusters. It then performs clustering on the dissimilarity matrix to obtain the final partition.

The MI approach uses MI to measure and quantify the statistical information shared between a pair of clusterings. It can automatically select the best clustering

method from several algorithms. Examples are as follows. A Genetic Algorithm (GA) clustering ensemble [16] uses a GA to obtain the best partition and the co-association function as the consensus function. It determines fitness function parameters based on co-association function values. The information theory based GA clustering ensemble [176] uses a GA to find a combined clustering by minimizing an information-theoretical criterion function. The generalized MI clustering ensemble [265] introduces a new consensus function using a generalized mutual information definition. The consensus function is related to the classical intraclass variance criterion.

The co-association based functions approach is also called the pair-wise approach. It uses a co-association matrix in the consensus step. Examples are as follows. Clusterfusion [144] first generates an agreement matrix with each cell containing the number of agreements amongst clustering methods and then uses the matrix to cluster data objects. Voting-k-Means [89] transforms data partitions into a co-association matrix with coherent association mappings. It then extracts underlying clusters from this matrix. Evidence accumulation-based clustering [90] maps data partitions created by each individual clustering into a new similarity matrix, based on voting. It then uses the single link algorithm to extract clusters from this matrix.

Finite mixture model approach assumes that the probability of assigning a label to a data object is based on a finite mixture model or that the labels are 'modeled as random variables drawn from a probability distribution described as a mixture of multivariate component densities' [270]. It obtains the consensus clustering result by solving a maximum likelihood estimation problem. Mixture model clustering ensemble [266] uses a probabilistic model of consensus based on a finite mixture of multinomial distributions in a space of clusterings. It finds a combined partition by solving the corresponding maximum likelihood problem with the EM algorithm.

The graph/hypergraph partitioning approach considers the combination problem as a graph or hypergraph partitioning problem. Methods taking this approach differ in how they build a (hyper)graph from the clusterings, as well as how they define the cuts on the graph to obtain the consensus partition [270]. Examples are as fol-

lows. METIS [141] is a multi-level graph partitioning system. It collapses vertices and edges of the graph, partitions the resulting coarsened graph, and then refines the partitions. SPEC (spectral graph partitioning algorithm) [202] tries to optimize the normalized cut criterion. It treats the rows of the largest eigenvalues matrix as multiple dimensional embeddings of the vertices of the graph and then uses k-means to cluster the embedded points. CSPA (Cluster based Similarity Partitioning Algorithm) [259] first creates a graph based on a co-association matrix, and then performs METIS clustering on the graph. HGPA (Hypergraph Partitioning Algorithm) [259] uses a hyperedge in a graph to represent each cluster. It then uses minimal cut algorithms such as HMETIS [139] to find good hypergraph partitions. MCLA (Meta Clustering Algorithm) [259] determines soft cluster membership values for each data object by using hyperedge collapsing operations. HBGF (Hybrid Bipartite Graph Formulation) [85] constructs a bipartite graph where data objects and clusters are both modeled as vertices. It later partitions the bipartite graph with an appropriate graph partitioning method.

Multi-objective Clustering:

Conventional clustering algorithms use a single clustering objective function only, which may not be appropriate for the diversities of the underlying data structures. Multi-objective clustering uses multiple clustering objective functions simultaneously. Such methods consider clustering as a multi-objective optimization problem [84].

Examples are as follows. FCPSO (Fuzzy Clustering-based Particle Swarm Optimization) [7] uses an external repository to save non-dominated particles during the search process and a fuzzy clustering technique to manage the size of the repository. It also uses a fuzzy-based iterative feedback mechanism to determine the compromised solution among conflicting objectives. Evolutionary Multiobjective Clustering [111] and MOCK (MultiObjective Clustering with automatic k-determination) [112] use an evolutionary approach to solve the multi-objective problem in clustering. They are based on a multi-objective evolutionary algorithm named PESA-II (Pareto Envelope-

based Selection Algorithm version 2) [56] to optimize two complementary clustering objectives. Multi-objective real coded genetic fuzzy clustering [199] aims to optimize multiple validity measures simultaneously. It encodes the cluster centers in its chromosomes while optimizing the fuzzy compactness within a cluster and fuzzy separation among clusters. EMO-CC (Evolutionary MultiObjective Conceptual Clustering) [231] combines evolutionary algorithms with multi-objective optimization techniques and relies on the NSGA-II multi-objective genetic algorithm [65]. It can discover less obvious but informative data associations.

2.3 Applications of Clustering in Cancer Subtyping

The recently-developed DNA microarray and sequencing technologies [52, 79, 168], which can measure the expression levels of tens of thousands of genes simultaneously, offer cancer researchers novel methods to investigate the pathology of cancers from a molecular angle. Under such a systematic framework, cancer types or subtypes can be identified through the corresponding gene expression profiles. Research on gene expression profile-based cancer type recognition has already attracted numerous efforts from a wide variety of research communities [190, 285]. Investigations on leukemia [104], lymphoma [8], colon cancer [9], cutaneous melanoma [30], bladder cancer [78], breast cancer [217], lung cancer [94], and others show very promising results. Supervised computational methods, such as multi-layer perceptrons [147], naive Bayes [164], support vector machines [166, 257], semi-supervised Ellipsoid ARTMAP [284], and k-Top Scoring Paris [260], have already been used in cancer diagnosis-oriented gene expression data analysis.

In this section, we consider the situation in which we do not have labels for the cancer samples. This assumption is reasonable with the requirement for discovering unknown and novel cancer types or subtypes. In this case unsupervised learning or cluster analysis [285] is required in order to explore the underlying structure of the obtained data and provide cancer researchers with meaningful insights into the

possible partitions of the samples.

One of the major challenges of microarray data analysis is the overwhelming number of measures of gene expression levels compared with the small number of samples, which is caused by factors such as sample collection and experiment cost. This problem is well known as the 'curse of dimensionality' in machine learning, which refers to the lack of data separation in high dimensional data space. When the dimensions are high, the distance from a data object to the nearest neighbor data object becomes indistinguishable compared with the majority of data objects [26].

There are multiple steps to obtain microarray data, due to several system or design issues, and each step may introduce noise. Noise can obscure or mislead the underlying biological meanings, which is an important reason why statistical tools are used to analyze microarray data, since they can take the noise or variations into account. The noise can come from five phases of data acquisition: microarray manufacturing, preparation of mRNA from biological samples, hybridization, scanning, and imaging [50]. And they can be classified into three major categories: biological - cells from different populations, tissues, conditions, etc. experimental - defects of the spotting equipment, different hybridization conditions and dyes, different methods to make the arrays, to culture the cells, to extract mRNA, etc. processing - errors related to numerical values collection such as fluorescence scanning, image analysis, and intensity readout [299].

mRNA profiling has demonstrated its effectiveness at subtyping various cancers. miRNA (short for MicroRNAs) profiling can be more accurate. Researchers have found links between misregulated miRNAs and the genes that are affected in various cancer subtypes [212]. miRNAs are small non-protein coding RNAs found in animals and plants. The first miRNAs were discovered and characterized in the early 1990s [162]. Since the early 2000s, miRNAs have been found to play multiple roles in negative regulation in cells. The first cancer found to be associated with miRNA deregulation and deletion was chronic lymphocytic leukemia [38]. Later on many miRNAs have been found to be related to more types of cancer [33, 41, 51, 126, 134,

138, 193, 194, 290]. Since multiple subtypes of a disease may have similar patterns within a single data type (*i.e.*, mRNA or miRNA), both data types can be used together to improve the accuracy of subtyping.

2.3.1 Clinical Applications

mRNA-based Applications

Golub used mRNA profiling of the expression of 6,817 genes in 72 leukemia samples as a test case for subtyping [104]. Using self-organizing maps (SOMs), leukemia samples were successfully grouped into the known subtypes of acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) without previous knowledge of these subtypes. The results showed the feasibility of using gene expression alone to classify cancer and suggested a general approach of classification for other types of cancer without using previous biological knowledge.

Alizadeh used mRNA profiling to study 128 microarray analyses that contain 1.8 million measurements of gene expression from 96 samples of normal and malignant lymphocytes [8]. Using the hierarchical clustering approach, two subtypes of diffuse large B-cell lymphoma (DLBCL) were identified: germinal center B-like DLBCL, which is diverse in gene expression patterns, and activated B-like DLBCL, which is distinct at the molecular level. Patients with germinal centre B-like DLBCL had a significantly better response to current therapy and overall survival than those with activated B-like DLBCL, which reflects tumor proliferation rate, different state of the tumor, and different host-patient response.

Armstrong applied clustering technique of PCA on mRNA profiling of 8700 genes from 72 leukemia samples and discovered mixed-lineage leukemia (MLL), a leukemia subtype that is distinct from both AML and conventional ALL [14]. MLL is characterized by the mixed-lineage leukemia gene's chromosomal translocation, and such patients have a decidedly poor prognosis and often have early relapse after chemotherapy. The discovery of MLL as a distinct subtype is important to therapeutic success as well, since molecular markers differentially expressed by MLL compared with both

ALL and AML immediately suggest new and different molecularly targeted treatment strategies for this treatment-resistant cancer subtype.

Bittner studied mRNA profiling of 6971 genes from 31 patients with malignant melanoma, for which there were no accepted histopathological, molecular, or immunohistochemical-marker defined subtypes [30]. Hierarchical clustering (agglomerative, average linkage) with Pearson correlation coefficients discovered two potential subtypes. With in vitro assay experiments, the subtypes were associated with different disease tissue invasion potential [182]. However, the patients in this study had uniformly poor prognosis, and future work is needed to analyze the clinical relevance of observed subtypes.

Perou analyzed mRNA profiles of 8102 genes of 65 breast tumor specimens using a hierarchical clustering approach [217]. Three subtypes were discovered in this clinically highly heterogeneous tumor: the previously known Erb-B2, and two others previously unknown, namely ER+(estrogen receptor-positive)/luminal-like and basal-like. Due to the limited number of tumor specimens in this study, statistically significant relationships between the discovered subtypes and clinical data are still to be uncovered.

Lapointe profiled mRNA gene expression of 26,260 genes in 62 primary prostate tumors and 9 lymph node metastases and identified three robust subtypes of prostate tumors using a two-way hierarchical clustering technique on 5153 genes based on distinct gene expression patterns [159]. Subtype I is the clinically least aggressive subgroup, subtype II is the second clinically aggressive subgroup, and subtype III is the most clinically aggressive subgroup, including most of the metastasis cases in this study. These tumor subtypes may provide a basis for improved prognostication and treatment decision.

Liang performed agglomerative hierarchical clustering on mRNA profiles of 1800 genes from 32 samples including Glioblastoma multiforme (GBM) and normal brain [167]. Two molecularly distinct subtypes of GBM were identified, and their expression showed an obvious difference in a group of genes correlated with survival. Such finding

may improve the accuracy of prognostic predictions and facilitate the development of optimized therapies for each subtype.

Laiho showed that mRNA profiling of 7928 genes from 37 colorectal carcinoma (CRC) samples separated serrated CRCs and conventional CRCs using hierarchical clustering [158]. This study was able to provide firm molecular evidence for a previously underrecognized route leading to CRC, which is a serrated neoplasia pathway. Much clinical and pathological evidence suggested that serrated CRCs may be more aggressive than conventional CRCs. Establishing serrated CRCs as a biologically distinct CRC subtype represents further discovering of the molecular subtypes of CRCs. In the long term, understanding the molecular basis of serrated CRCs may contribute to the development of treatment options specifically for this tumor subtype.

Wilkerson detected four lung squamous cell carcinoma (SCC) subtypes from mRNA expression data totaling 2307 genes from 382 SCC patients using Consensus Clustering in the ConsensusClusterPlus software package by Bioconductor [197, 278, 277]. The four lung SCC subtypes are: primitive, classical, secretory, and basal. These subtypes were associated with tumor differentiation as well as patient gender. The primitive subtype had the shortest survival and can be used as an independent predictor for survival outcome. The expression profiles of the four subtypes showed different biological processes which may suggest different pharmacologic interventions.

Lei et al. identified 3 major subtypes among mRNA expression data of 35 genes from 248 gastric tumors using a robust method of unsupervised clustering, consensus hierarchical clustering with iterative feature selection [163]. The 3 subtypes of gastric adenocarcinoma are: proliferative, metabolic, and mesenchymal. These subtypes have differences in molecular and genetic features, and respond differently to therapy. Thus, such subtyping may be helpful in selecting specific and appropriate treatment approaches for patients.

microRNA-based Applications

Lu et al. performed computational analyses on 217 miRNAs from 334 mammalian samples, including multiple human cancers [175]. Hierarchical clustering with average linkage and Pearson correlation was performed. Over miRNA profiles of 73 ALL samples, three groups were separated: BCR/ABL-positive samples and TEL/AML1 samples; T-cell ALL samples; and MLL samples. Subtyping results based on miRNA profiles on ALL samples and other tumor samples showed higher accuracy when compared with mRNA profiles. These discoveries demonstrate that using miRNA profiling for cancer diagnosis is very promising.

Blenkiron reported the clustering analysis of miRNA expression in primary human breast tumors [31]. Hierarchical clustering with average linkage and Pearson correlation were used: ER- and ER+ tumors were recovered in over 137 miRNAs in 93 primary tumors samples; basal-like, HER2+, luminal A, luminal B or normal-like were recovered in over 38 miRNAs in 51 tumor samples; luminal A or luminal B tumors were recovered in over 9 miRNAs in 24 tumor samples. This study is among the first integrated analysis using miRNA expression, mRNA expression and genomic changes in human breast cancer. Furthermore, it demonstrates that miRNA expression profiling has the potential to effectively classify breast cancer into prognostic molecular subtypes.

Mattie analyzed miRNA profiling of 20 different breast cancer samples in three common subtypes: ErbB2+/ ER-, ErbB2+/ER+, and ErbB2-/ER+ [187]. Hierarchical clustering identified these clinically relevant subtypes based on their miRNA expression patterns. The ErbB2+/ER+ subtype is a clinically troublesome subtype and appears to be more resistant to all kinds of endocrine therapy [63]. Successfully identifying the ErbB2+/ER+ subtype based on miRNA profiling is of substantial interest since mRNA profiling studies had not previously been able to identify it.

Porkka studied the expression of 319 human miRNAs in samples from prostate cancer cell lines, prostate cancer xenografts and clinical prostate tissues [223]. Hierarchical clustering with average linkage separated the 9 prostate carcinoma tissue

samples into two groups that quite accurately correspond to clinical stage based subtypes: hormone-naive subtype and hormone-refractory subtype. Such results indicate that miRNAs profiling has the potential to become a novel diagnostic and prognostic tool for prostate cancer.

Oberg examined the expression of 735 miRNAs in 52 normal and 263 colon tumor samples [211]. There were three clinical subtypes in the tumor samples: 41 adenomas, 158 pMMR carcinomas and 64 dMMR carcinomas. Hierarchical clustering with average linkage and Pearson's dissimilarity matrix demonstrated that normal colon tissue and the three tumor subtypes were all clearly separable. It is the first report to show global miRNA (instead of only a few selected miRNAs) expression differences can be used for colon tumor subtype diagnosis.

Yang analyzed 219 miRNA-associated genes from 459 ovarian carcinoma (OvCa) samples [292]. Consensus k-means clustering identified two clusters. One of the two clusters contained 172 OvCa cases and formed a tight cluster with higher expression values of the miRNA-associated genes. The majority of patients in this cluster had advanced stage OvCa and significantly shorter overall survival durations than patients the other cluster.

2.3.2 Computational Experiments

mRNA-based Experiments

Xing and Karp presented CLIFF (CLustering via Iterative Feature Filtering) [282] and tested it with mRNA profiles of 72 leukemia samples and 7130 genes [104]. CLIFF is based on the 'normalized cut' concept and iterates between sample partitioning and feature filtering until converging into an appropriate partition of the leukemia samples and a set of informative genes. The result produced by CLIFF had high agreement to the original expert labeling of the leukemia data set. Its final partition had two clusters. One of the clusters contains 44 ALL samples, and the other cluster contains 25 AML samples and 3 ALL samples.

Tang and Zhang proposed IPD (Iterative Pattern-Discovery) based on iterative

sample clustering and irrelevant gene pruning [262], and tested it on mRNA profiles of over 7129 genes in 72 leukemia patient samples [104]. During the initial partition phase, conventional clustering methods k-means or SOM was used to group samples and genes into exclusive smaller groups. Based on Rand Index values, the clustering results obtained by IPD approaches are consistently better than the results obtained by applying k-means or SOM directly.

Getz *et al.* proposed CTWC [98] (covered in Section 2.2.7) and applied it to a leukemia mRNA data set [104] and a colon cancer mRNA data set [9]. The leukemia data set contains 72 samples; 47 samples are ALL, and the other 25 samples are AML. The original set contained 6,817 genes. 1,753 genes were selected for the CTWC experiment. In two iterations, 49 stable gene clusters and 35 stable sample clusters were obtained. One of the gene clusters contained 60 genes, and when used as the feature set CTWC was able to separate the samples into AML/ALL clusters. The colon cancer data set contained 40 tumor samples and 22 normal samples. The original set contained 6,500 genes. 2,000 genes were chosen for CTWC experiment. In two iterations, 97 stable gene clusters and 76 stable sample clusters were obtained by CTWC. Four of the gene clusters can partition the samples into normal/tumor clusters.

Cheng and Church proposed an efficient biclustering algorithm [47] and applied it to diffuse large B-cell lymphoma mRNA profiles containing 4026 genes and 96 conditions [8]. The algorithm is based on multiple row/column addition/deletions and successively extracts biclusters from the raw data matrix until a pre-specified number of clusters has been reached. In comparison with the results from hierarchical clustering used by Alizadeh et al., the first 100 biclusters discovered by Cheng's biclustering had only 10 conditions/genes exclusively from one or the other primary cluster from hierarchical clustering.

Iam-on *et al.* presented LCE (Link-based Cluster Ensemble) [124] and applied it to several mRNA profile data sets including: leukemia1 [104] (1,877 genes and 72 samples), leukemia2 [104] (1,877 genes and 72 samples), leukemia3 [14] (2,194 genes and

72 samples), brain tumor [210] (1,377 genes and 50 samples), central nervous system (CNS) [222] (1,379 genes and 42 samples), and hepatocellular carcinoma (HCC) [44] (85 genes and 180 samples). LCE incorporates relations within an ensemble and associations among clusters to improve clustering results. Based on average validity measure over three validity indices (Classification Accuracy (CA), Normalized Mutual Information, and Adjusted Rand Index) on the clustering results, LCE regularly performs better than other clustering methods including MULTI-K [148], consensus clustering with hierarchical clustering [197], graph-based consensus clustering [295], Cluster based Similarity Partitioning Algorithm [259], Hyper-Graph Partitioning Algorithm [259], Meta-Clustering Algorithm [259], Hybrid Bipartite Graph Formulation [86], k-means, single-linkage, complete-linkage, and average-linkage. Based on the CA validity index, LCE achieved over 74% accuracy on leukemia1, over 70% accuracy on leukemia2, over 83% accuracy on leukemia3, over 61% accuracy on brain tumor, over 63% accuracy on CNS, and over 84% accuracy on HCC.

microRNA-based Experiments

Lock and Dunson proposed Bayesian Consensus Clustering (BCC) [172] and tested it with miRNA profiles of 348 breast cancer samples and 423 miRNAs [149]. This approach is a flexible and computationally scalable Bayesian framework, which estimates the consensus clustering and the base clusterings at the same time. BCC clustering results vs. TCGA identified comprehensive subtypes matching matrix show that the two partitions have a significant but weak association.

Li *et al.* proposed a subtyping method using the CTWC algorithm and Super-Paramagnetic Clustering (SPC) [165] and tested it with the miRNA profiles of 71 breast cancer patients and 13 miRNAs [31]. This method iteratively partitioned the sample and feature space using the two-way super-paramagnetic clustering technique and identifies the final optimal miRNA clusters. Using a subset of the miRNAs as the feature set, the five subtypes previously classified by mRNA expression profiling [31] were identified successfully by CTWC. The clinical significance of the identified sub-

types were verified using Kaplan-Meier survival analysis [151].

2.4 Challenges

Despite many examples of successful applications of cluster analysis, there still remain many challenges due to the existence of many inherent uncertain factors. The following fundamental challenges in clustering are relevant even today [129]: a) definition of a cluster, b) selection of features, c) normalization of the data, d) outlier detection, e) definition of pair-wise similarity, f) number of clusters, g) selection of clustering method, h) existence of clustering tendency, and i) validity of the clusters.

Advances in expression profiling technology and decreasing costs are making gene expression data increasingly available and affordable. Research has shown that classifying cancers using gene expression can discover previously undetected and clinically significant subtypes of cancer [269]. However, there are still many challenges.

2.4.1 Clinical Challenges

Complexities in cancers and cancer subtypes Cancers and cancer subtypes are complicated diseases. Especially for most solid tumors, many different cell types are involved in a tumor, and tumor cells themselves are morphologically and genetically diverse [218]. These features may make the conventional clustering approaches problematic or inadequate, so novel clustering approaches are needed to address such complexities.

Experimental issues Gene expression studies require careful experimental design to avoid experimental errors. This is especially important for studying solid tumors. For example, biopsy specimens might have different proportions of surrounding stromal cells, which may cause clustering results reflecting the stromal contamination, rather than the underlying tumor. So, additional techniques are needed to improve such problems. Microscopic examination of tumor samples to make sure that the tumor cells are comparable and purified is helpful, as well as computational analysis methods that can exclude surrounding stromal cells.

2.4.2 Computational Challenges

Curse of Dimensionality Gene expression data sets generally contain small numbers of samples (in tens) and large numbers of genes (in hundreds or thousands). Most conventional clustering techniques need a large number of samples and a small number of variables to achieve robust performance. Approaches are needed to improve the clustering results on such sparse data sets.

Noise There are multiple sources of noise introduced in microarray experiments, including varying cellular composition among tumors, genetic heterogeneity within tumors due to selection and genomic instability, differences in sample preparation, nonspecific cross-hybridization of probes, and differences between individual microarrays. In general, biologic variation is the major source of variation in gene expression experiments. The noise may obscure clustering results, especially in those approaches based on distance functions. Techniques are needed to improve clustering approaches so that they are more robust to noise.

Number of subtypes For subtyping in clinical studies, the number of subtypes are unknown or uncertain. However, in many clustering algorithms this number needs to be specified by the user, so techniques are needed to estimate or infer the number from the data. Algorithms are needed to identify the number of clusters [208].

Clinical or biologic meanings Gene expression data sets are complex and may contain hundreds or thousands genes. In a complex data set, many different relationships and patterns are possible [227]. The patterns discovered by clustering may not necessarily be clinically or biologically meaningful. Techniques are needed to uncover and identify clusters of clinical or biologic interest.

Statistical significance An important issue with any analytic approach is the statistical significance of observed correlations. A typical microarray experiment produces expression data for thousands of genes from a relatively small number of samples, thus gene-cluster correlations can be identified by chance alone. Techniques are needed to determine the statistical significance, such as permutation testing.

Knowledge integration Knowledge is obtained from multiple test techniques: some from conventional tests and some from molecular diagnostic tests. However a single recommendation is needed for the oncologist to treat the patient. Approaches are needed to integrate these knowledge items to produce an improved single recommendation.

Algorithm selection There are a large number of clustering algorithms available that may be used for clustering gene expression data, however there is no single best algorithm that performs best in all aspects. Selecting the most appropriate algorithm for a given gene expression data set and a given analysis goal is critical in success application. Without automatic selection of the most appropriate algorithm, researchers usually select a few promising clustering algorithms and compare their results. Approaches are needed to improve algorithm selection.

Other challenges Besides the above challenges, researchers are also facing the following challenges: time variation during specimen preparation, integration of data sets created by different laboratories using different technologies, overlapping clusters, presence of irrelevant attributes, and lack of prior knowledge.

2.5 Discussion

In this chapter we reviewed classical and state of the art clustering algorithms in the communities of computer science, machine learning, statistics, etc. We also reviewed historic and state of the art cancer subtyping techniques. Clustering algorithms that have been applied to mRNA or miRNA expression data based cancer subtyping with promising results, and challenges associated with molecular cancer subtyping were presented as well.

However, given the many choices of available clustering algorithms, there is no single algorithm that performs best in every validation matrix. The performance of a given clustering algorithm and a validation matrix is dependent on data characteristics and the application [133].

Different granularities Users often desire different cluster granularity for dif-

ferent subsets of data. For example, users may prefer small and tight clusters for some genes but need only coarse data structure for other genes. However, most existing clustering algorithms provide the same cluster granularity for all genes. It would be more helpful for them to provide a flexible representation of the data cluster structure and let the user to find the answer based on several different granularity requirements.

Handling high dimensional / low sample data Although many clustering techniques have been used for gene expression data, most of these techniques perform well only on data with a large number of samples and a small number of dimensions. Cancer gene expression data sets generally contain a small number of samples with a large number of dimensions or genes, since many human cancer studies use costly or rare clinical specimens and are difficult to repeat. Future advances will require improved clustering techniques better adapting to this type of data.

Easy to use software In order to make routine clinical use of clustering tools a reality among medical and biological professionals, the software needs to be easy to use. For example, the software should be able to determine the number of clusters automatically based on the data properties; the software should avoid needing other user-specific parameters or should provide effective guidance to determine those parameters; the software should provide good visualization and domain-specific interpretation for the clustering results; the software should be able to extract useful and relevant information from the data to solve users' problems.

Robustness Noise can be introduced in every step of the microarray experiments due to the nature of microarray technology. It is not realistic to count on the data to be 'pure and uncontaminated'. Noise and outliers can be present in the data during measurement, storage, and processing. The clustering algorithm should be able to better detect and remove noise and outliers, or not be affected by them.

Arbitrary cluster shapes Many existing clustering algorithms form clusters with regular shapes, such as hyper-spheres or hyper-rectangles. Gene expression data generally have complex underlying data structures and are not always regular cluster shapes. Cluster algorithms should be able to better detect arbitrary natural cluster

shapes rather than confine the clusters to some particular shape.

Chapter 3

Improved Fuzzy Cluster Ensemble Methodology

3.1 Introduction

Many clustering methods have been designed and applied to cancer gene expression data for the purpose of cancer classification. They aim to improve therapeutic results by diagnosing cancer types or subtypes with improved accuracy in comparison with traditional methods such as histopathology or immunohistochemistry.

A common and exploratory analysis is to perform clustering on the cancer or patient samples (tissues). Such kind of analysis was first carried out in late 1990s with promising results. In addition, bioinformaticians have proposed novel clustering methods that take intrinsic characteristics of gene expression data into account, such as noise and high-dimensionality, to improve the clustering results. However, different algorithms (or even the same algorithm with different parameters) often provide distinct clusterings. As a result, it is extremely difficult for users to decide which algorithm and parameters will be optimal for a given set of data set for a particular task. There is no single clustering algorithm that can perform the best for all data sets [156], and discovering all types of cluster shapes and structures presented in data is impossible for any known clustering algorithm [75].

Cluster ensembles have recently emerged as simple and effective methods for improving the robustness and accuracy of clustering results. Cluster ensemble can perform many algorithms on a data set, and integrate the results to find the best

clustering.

3.2 Related Work

Clinical researchers usually use simple traditional clustering methods such as hierarchical [254], K-Means [238], and SOM [104] for cancer gene expression data cluster analysis. Such traditional methods have much better availability in standard software packages and are easy to implement.

Novel clustering methods have been proposed by bioinformaticians to improve the clustering results on gene expression data to address its intrinsic characteristics including noisy and high dimensional, such as Non-negative Matrix Factorization (NMF) method [36]. Such new methods are not getting enough attention from clinical researchers as they may require particular programming environments or more user-specified parameters, which is difficult for non-expert users.

Cluster ensembles combine multiple clustering decisions from base clusterings or ensemble members. There are two main steps in a clustering ensemble: generation step and consensus step.

In generation step, cluster ensemble methods use a variety of approaches to obtain diversity in base clusterings. Four ensemble generation methods have been commonly used: a) using a single clustering algorithm with different initializations [148], b) using multiple clustering algorithms [157], c) using different subsets of genes [15], and d) using data sampling techniques [76].

In consensus step, cluster ensemble methods use a variety of consensus functions to combine base clusterings. Four ensemble consensus methods have been commonly used: a) using pairwise similarity-based consensus function [148], b) using graph-based consensus function [259], c) using mutual information-based consensus function [265], and d) using voting based consensus function [76].

Noise problem remains particularly challenging in clustering applications, even if there are many pre-processing techniques such as logarithmic transformation or standardization. For bioinformatics applications, noise can make it difficult to detect

the true clusters and obscure or mislead the underlying biological meanings.

We present a novel clustering algorithm Improved Fuzzy Cluster Ensemble (IFCE) to improve robustness against noise.

3.3 Noise Robustness Problem

Noise is the opposite of true signal data objects in a data set, and is meaningless additional information. Noise can be present in the data during measurement, storage, processing, and collecting. It is not practical to count on the data to be free of noise. Figure 3.1 shows data (noise + signal) against signal in a very low signal-to-noise data set. Figure 3.2 shows data (noise + signal) against noise in a very low signal-to-noise

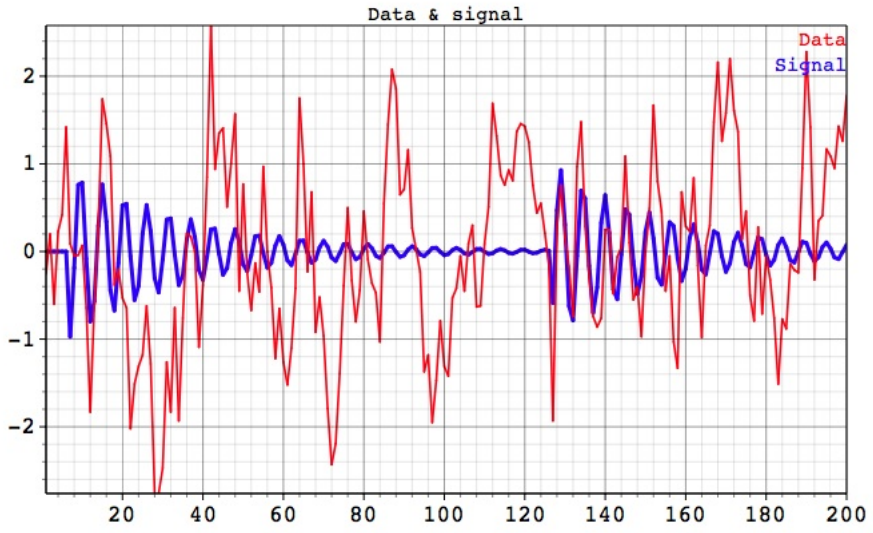

Figure 3.1: Data and signal [256]

data set.

Noise problem remains challenging in clustering applications, even if there are many pre-processing techniques such as logarithmic transformation or standardization. For bioinformatic applications, noise can make it difficult detect the true clusters and obscure or mislead the underlying biological meanings. The noise may obscure clustering or mislead results especially in those approaches based on distance func-

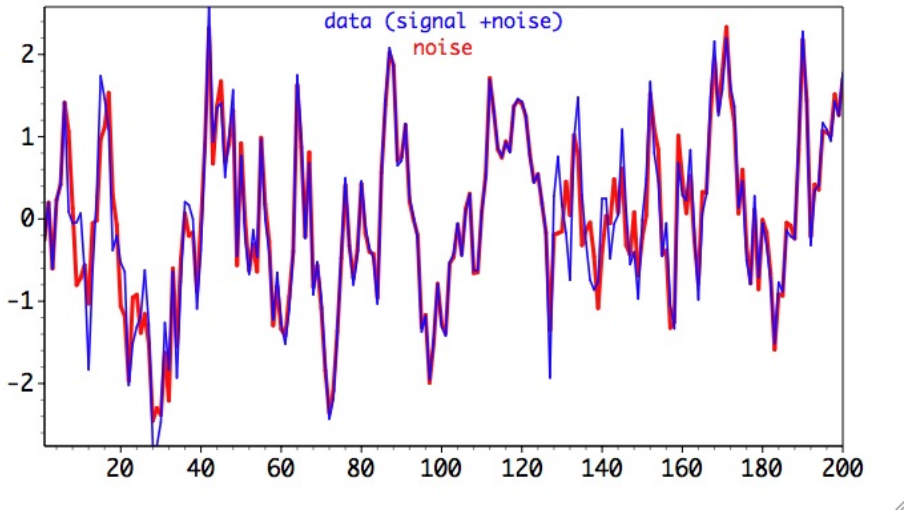

Figure 3.2: Data and noise [255]

tions. An ideal clustering algorithm should be able to better detect and remove noise or not be affected by them [25].

Existing clustering methods that use a distance (e.g. Euclidean distance) as similarity measure, are not immune to the noise and can cause misclassification. Distance function combines the feature noises into a single similarity value. and Significant noise on some feature(s) cause the clustering results misleading [174]. Trimming methods [60, 204, 95] have been proposed to improve robustness. Such methods discard a predefined noise fraction (e.g. α) of an input data set before applying a clustering algorithm, which can be effective but suffer from exponential computational complexities. Since noise or outlier data objects can be considered as separate clusters, methods [204] have been proposed to increase the number of clusters when clustering noisy data. However, such methods are not proven for noise robustness guarantees. [25]

With our proposed algorithm IFCE, new weighted fuzzy techniques are employed to increase its robustness against noise.

3.4 Fuzzy Set Theory

Most of traditional methods for modeling, reasoning, and computing are crisp, deterministic, and precise. This means they are dichotomous, yes-or-no instead of more-or-less. [302] They use traditional Boolean logic, which takes only binary or dual values of either true or false that are usually denoted 1 or 0 respectively. No values in between is accepted. In traditional set theory, an element either belongs to a set of not.

However, reality is not crisp or certain. The problems in the real world are not always yes-or-no type. Real situations are often more-or-less or vague. So, traditional Boolean methods mostly not applicable well. Many theories have been developed to model uncertainties in reality. For a long time, probability theory and statistics have been the predominant one. However it is based on certain assumptions as traditional theories, which can be different than reality [302, 136].

Fuzzy set theory is also one of those theories and was introduced in 1965. It uses fuzzy logic, which is a many-valued logic and takes the values as any real numbers between 0 and 1. It was initially intended to be an extension of traditional Boolean logic and traditional set theory. It provides a natural way of dealing with problems that do not have sharply defined criteria of set membership. It describes mathematically the vagueness or imprecision. It is strict mathematical method in which vague situations can be precisely and rigorously studied. Although there is imprecision, humans can still make sensible decisions. The interests of this theory grew slowly till its first successful practical applications in fuzzy control systems in later 1970s and further many successful applications in 1980s. Ever since, it has been developed to be a powerful method and can often model reality better than traditional theories. It was expected to have the potential of a wider scope of applicability, particularly in the fields of pattern classification etc. [302, 136] Figure 3.3 shows an example of fuzzy logic. Figure 3.4 shows examples of fuzzy logic applications.

Gene expression data contains imprecise information. Crisp or hard clustering

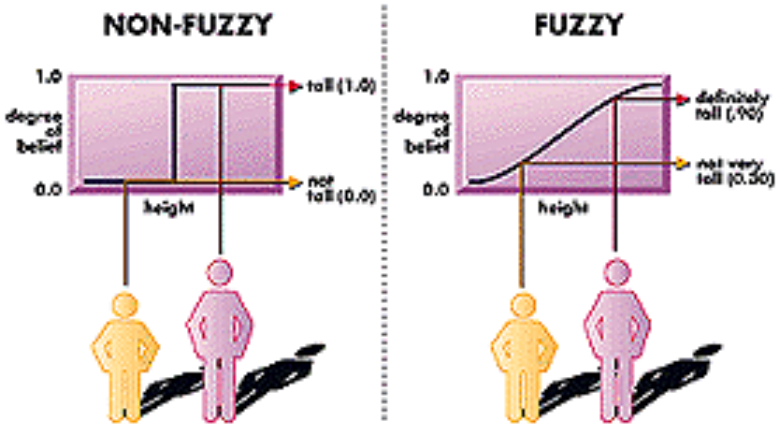

Figure 3.3: Fuzzy logic [17]

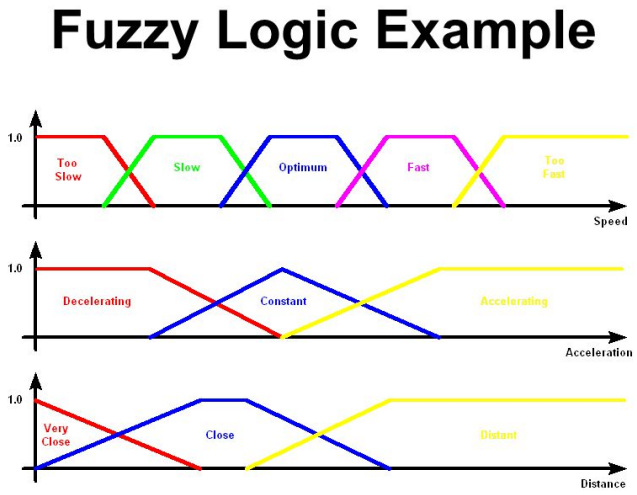

Figure 3.4: Fuzzy logic examples [245]

methods such as K-means and SOM are not suited to the analysis of such data because the clusters of genes frequently overlap. Fuzzy set theory has many advantages in dealing with data containing imprecision. Fuzzy clustering approaches use fuzzy set theory which takes this imprecision into consideration in analyzing gene expression data [67].

With fuzzy clustering, a cluster is viewed as a fuzzy set. Thus, each feature

vector has a membership value with each cluster, which is the degree of belonging to each cluster. Figure 3.5 shows an example of fuzzy membership.

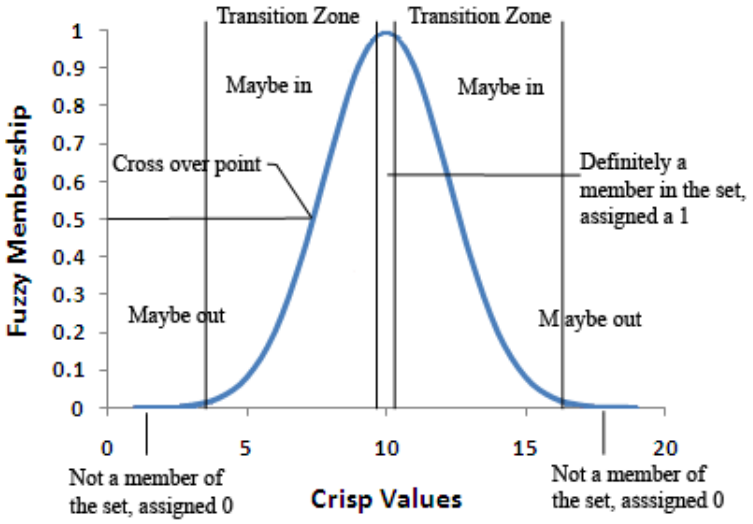

Figure 3.5: Fuzzy membership [224]

A membership function is a function that maps each point in the input space to a fuzzy membership value between 0 and 1 representing degree of membership. For a input point, the value 0 means it is not a member of the fuzzy set. The value 1 means it is fully a member. The values between 0 and 1 means it is a fuzzy member and belongs to the fuzzy set only partially. There are many fuzzy membership functions. Some of them are formed using straight lines such as the triangular membership function and the trapezoidal membership function. They are simplest. Gaussian membership functions are built on Gaussian distribution curve. The generalized bell membership function has one more parameter than the Gaussian membership function that defines its slope. These two achieve smoothness, but they are unable to specify asymmetric membership. Sigmoidal membership function is either open left or right, suitable for asymmetric membership. Polynomial membership functions are based on various Polynomial curves. Left-Right or L-R membership function uses both a left function and a right function which are monotonically decreasing functions. Also, there are 2-D membership function and composite of non-composite membership functions. [130, 181, 232, 229] Common fuzzy membership functions are presented in Table 3.1.

MF&Reference	Definition	Notes
Triangular [130, 229]	$\text{triangle}(x;a,b,c) = \begin{cases} 0, & x \le a. \\ \frac{x-a}{b-a}, & a \le x \le b. \\ \frac{c-x}{c-b}, & b \le x \le c. \\ 0, & c \le x. \end{cases}$	Simple formulas and computational efficiency. Not smooth at the corner points.
Trapezoidal [130, 229]	$\text{trapezoid}(x;a,b,c,d) = \begin{cases} 0, & x \le a. \\ \frac{x-a}{b-a}, & a \le x \le b. \\ 1, & b \le x \le c. \\ \frac{d-x}{d-c}, & c \le x \le d. \\ 0, & d \le x. \end{cases}$	Simple formulas and computational efficiency. Not smooth at the corner points.
Gaussian [130, 229]	$\text{trapezoid}(x;a,b,c,d) = \begin{cases} 0, & x \le a. \\ \frac{x-a}{b-a}, & a \le x \le b. \\ 1, & b \le x \le c. \\ \frac{d-x}{d-c}, & c \le x \le d. \\ 0, & d \le x. \end{cases}$	Smoothness and concise notation. Unable to specify asymmetry.
Generalized Bell [130, 229]	$\text{bell}(x;a,b,c) = \frac{1}{1+\left\lvert\frac{x-c}{a}\right\rvert^{2b}},$	Smoothness and concise notation. Unable to specify asymmetry.
Sigmoidal [130, 229]	$\text{sig}(x;a,c) = \frac{1}{1+\exp[-a(x-c)]},$	Asymmetric: open right or left, or close.
Left-Right(L-R) [130, 229]	$\text{LR}(x;c,\alpha,\beta) = \begin{cases} F_L\left(\frac{c-x}{\alpha}\right), & x \le c. \\ F_R\left(\frac{x-c}{\beta}\right), & x \ge c, \end{cases}$	Extremely flexible. Unnecessary complexity.

Table 3.1: Fuzzy membership functions

Figure 3.6 shows examples of four classes of parameterized MFs: (a) triangle (x; 20, 60, 80); (b) trapezoid (x; 10, 20, 60, 95); (c) Gaussian (x; 50, 20); (d) bell (x;

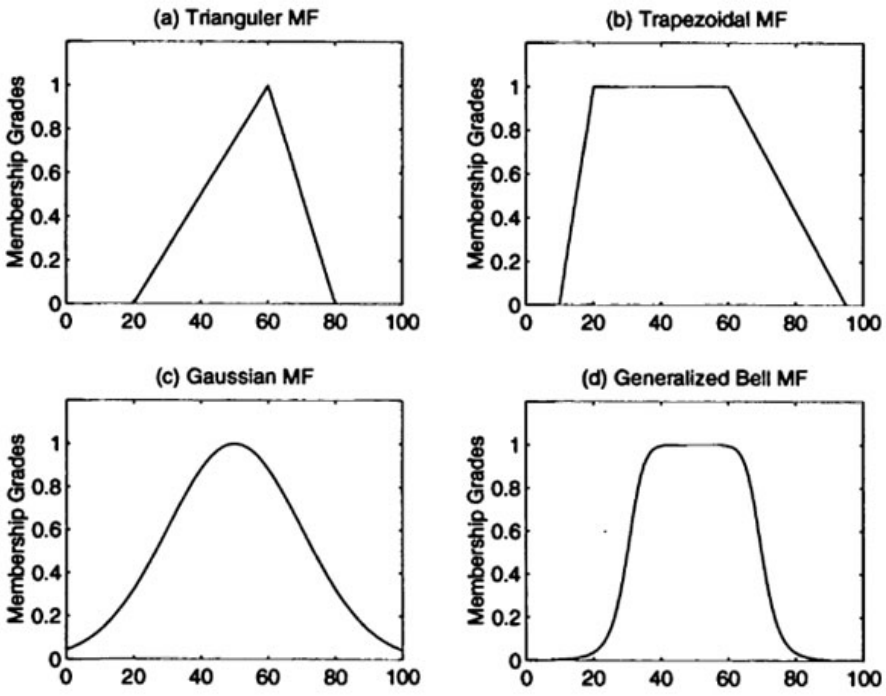

Figure 3.6: Parameterized fuzzy membership functions [229]

20, 4, 50) Figure 3.7 shows close and asymmetric MFs based on sigmoid functions: 1(a) shows two sigmoid functions y1 = sig(x; 1, -5) and y2 = sig(x; 2, 5); a close and asymmetric MF can be obtained by taking their difference $|y1 - y2|$, as shown in 2(b). 2(c) shows an additional sigmoid MF defined as y3 = sig(x; -2, 5); another way to form a close and asymmetric MF is to take their product y1y3, as shown in 2(d). Figure 3.8 illustrates two Left-Right(L-R) MFs specified by LR(x; 65, 60, 10) and LR(x; 25, 10, 40) [229].

3.5 Improved Fuzzy Clustering Algorithm

3.5.1 Description

The base clustering of IFCE is the Improved Fuzzy Clustering (IFC) algorithm. IFC is described below.

In IFC, it first generates a large number (twice of the total existing feature vectors) seed cluster centers in the feature space. It then eliminates those that are

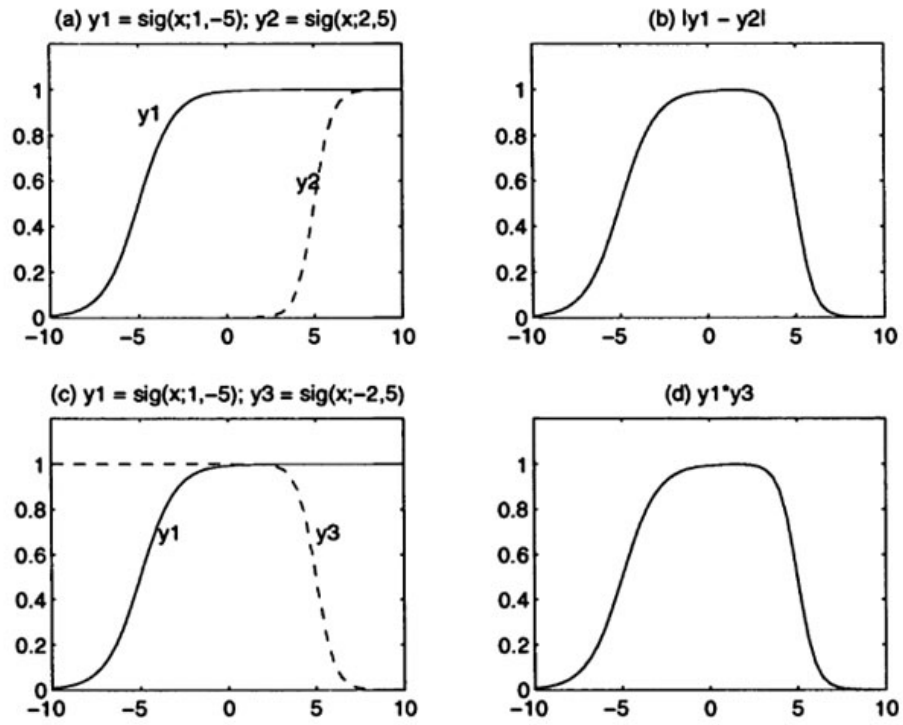

Figure 3.7: Sigmoid fuzzy membership functions [229]

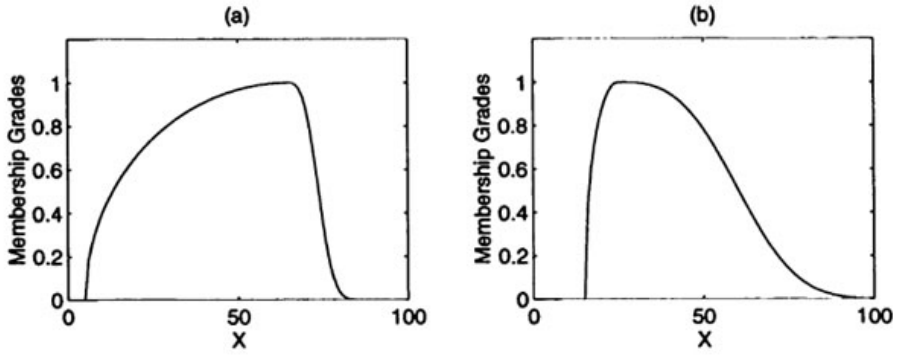

Figure 3.8: Left-Right(L-R) fuzzy membership functions [229]

too close to obtain a reduced but uniformly distributed set of initial seeds. For elimination, it uses the average distance between centers as a threshold, so half of the seeds are eliminated [174]. For a clustering problem, the number of clusters of a data set is often unknown. A general method starts with a large number of seeds (initial cluster centers) which include all the existing feature vectors and some randomly generated ones in the feature space [174].

After IFC obtains the initial cluster centers, it applies K-Means clustering (together, it is called Improved K-Means [173]). The K-means algorithm is used to partition a given set of observations into a predefined amount of K clusters. The algorithm as described by [178] starts with a random set of K center-points (μ). During each update step, all observations x are assigned to their nearest center-point (see equation 3.1). In the standard algorithm, only one assignment to one center is possible. If multiple centers have the same distance to the observation, a random one would be chosen. Traditional K-Means is given by the following two equations:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - \mu_i^{(t)} \right\|^2 \leq \left\| x_p - \mu_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \tag{3.1}$$

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{3.2}$$

IFC uses fuzzy clustering techniques. In non-fuzzy clustering (also called crisp clustering or hard clustering), data is divided into distinct clusters and each data object can only belong to exactly one cluster. In fuzzy clustering, each data object can potentially belong to multiple clusters. It uses membership grades to indicate the degree to which data objects belong to each cluster. Data objects on the edge of a cluster has lower membership grades, and is in the cluster to a lesser degree than data objects in the center of cluster with higher membership grades. Figure 3.9 shows Gaussian fuzzy set membership function. Gaussian fuzzy membership function

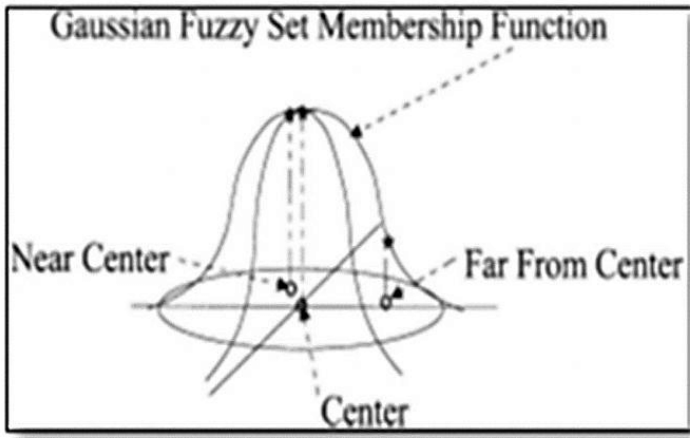

Figure 3.9: Gaussian fuzzy set membership function [3]

is given by the equation:

$$f(x\|\mu,\delta^2) = (\frac{1}{\sqrt{2\pi\delta^2}})exp[-\frac{(x_(p)-\mu)^2}{(2\delta^2)}] \tag{3.3}$$

μ is the mean or expectation of the distribution, δ is the standard deviation, and δ^2 is the variance. Fuzzy C-Means clustering is given by the equation [174]:

$$J(w_{qk},c^k) = \sum_{q=1}^{Q}\sum_{k=1}^{K}(w_{qk})^m||x^q-c^k||^2 \tag{3.4}$$

$$w_{qk} = \frac{(\frac{1}{(||x^q-c^r||)^2})^{\frac{1}{p-1}}}{\sum_{r=1}^{K}(\frac{1}{(||x^q-c^r||)^2})^{\frac{1}{p-1}}}, k=1,...,K, q=1,...,Q, \tag{3.5}$$

$$c^k = \sum_{q=1}^{Q} W_{qk}x^q, k=1,...,K; \tag{3.6}$$

$$W_{qk} = \frac{w_{qk}}{\sum_{r=1}^{Q} w_{rk}}, q=1,...,Q \tag{3.7}$$

IFC uses the modified weighted fuzzy expected value (MWFEV) method for computing the cluster centers [174]. Figure 3.10 shows Modified Weighted Fuzzy Expected Value. The Modified Weighted Fuzzy Expected Value (MFWEV) is given

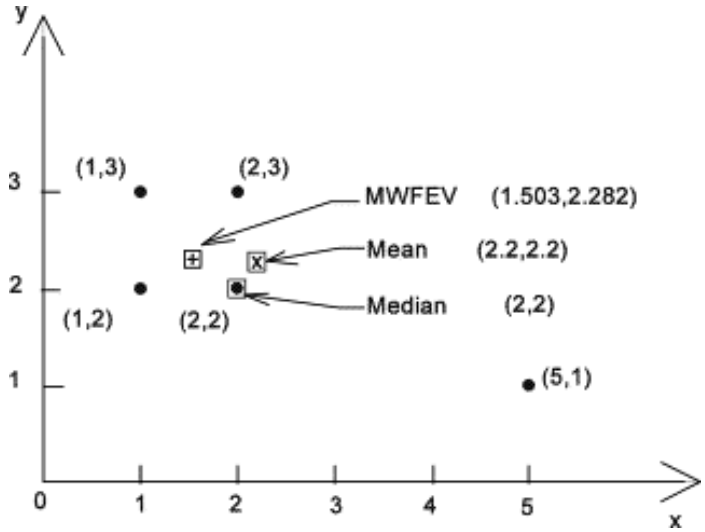

Figure 3.10: Modified weighted fuzzy expected value [174]

by the equation [174]:

$$\vec{\mu}^{(r+1)} = \sum_{p=1}^{P}\alpha_p^{(r)}x_p \tag{3.8}$$

Where,

$$\alpha_p^{(r)} = \frac{exp[-\frac{(x_(p)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]}{\sum_{m=1}^{P} exp[-\frac{(x_(m)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]} \tag{3.9}$$

$$(\delta^2)^{(r+1)} = \sum_{p=1}^{P} \alpha_p^{(r)} (x_p - \vec{\mu}^{(r)})^2 \tag{3.10}$$

After a number of IFC clustering iterations, it converges to many relatively small clusters ready for merging. In order to produce more natural shapes clusters as the results instead of forcing them into normed balls due to using the distance function, IFC merges the closest clusters until the Xie-Beni validity measure does not decrease anymore or until the number of clusters is reduced to two. It finds the two clusters with the minimum distance between their centers, calculates a new center with the average of the two centers. It then reduces the number of clusters by one accordingly [174].

Cluster merging [154] was proposed as a way to select the number of clusters. After the data set is partitioned into a relatively large number of clusters, similar clusters are merged based on a given criterion until no more clusters can be merged. This way, the number of clusters is reduced dynamically. There are various methods for cluster merging, including a compatible cluster merging method for clustering [154, 83], and the fuzzy inclusion similarity measure based method for an extended FCM algorithm [143]. In addition, some methods merge two clusters that has the largest pairwise linkage among all pairs of clusters. They use linkage functions such as the single linkage, the complete linkage, the average linkage, and the MinMax linkage [71]. Similarity-Driven cluster merging method for unsupervised fuzzy clustering was also proposed [283]

The Xie-Beni validity measure [281] measures the compactness and separation of the clustering results. It is the ratio of compactness-to-separation and defined by

$$XB = \frac{(\frac{1}{K}) \sum_{k=1}^{K} \delta_k^2}{D_{min}^2} \tag{3.11}$$

Where,

$$\delta_k^2 = \sum_{q=1}^{Q} w_{qk} ||x^q - c^k||^2, k = 1, ..., K. \tag{3.12}$$

$D_(min)$ is the minimum distance between cluster centers, and bigger value means greater separation. Each δ_k^2 is a fuzzy weighted mean-square error for the kth cluster, and smaller value means more compact clusters. Thus, a lower value of Xie-Beni means more compactness and greater separation [174].

Figure 3.11 shows an example of cluster merging.

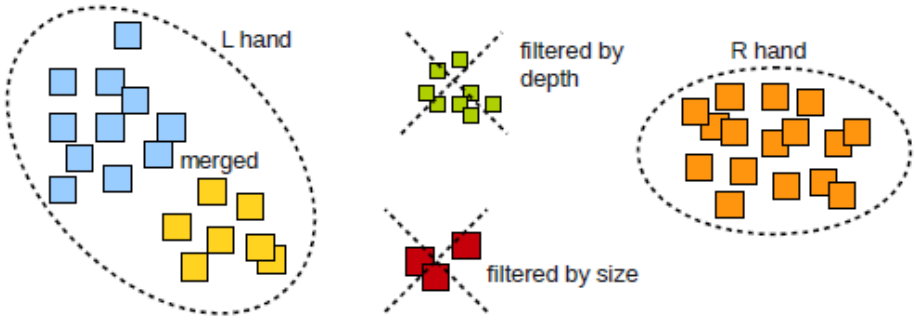

Figure 3.11: An example of cluster merging [137]

3.5.2 High Level Algorithm and Flowchart

The high level of IFC algorithm is given below, where $S = x(^q) : q = 1, ..., Q$ are input feature vectors, and $x(^q) = (x_1(^q), ..., x_N(^q))$; $c(^k) : k = 1, ..., K$ are cluster centers [174].

Step 1: Generate a large number of 2Q of seed cluster centers $c(^k) : k = 1, ..., 2Q$ and eliminate cluster centers that are too close to another to obtain a reduced set of K initial seed cluster centers.

Step 2: Assign each of the Q feature vectors $x(^q)$ to a cluster center $c(^k)$ based on minimum Euclidean distance via clust[q] = k to form K clusters.

Step 3: Eliminate all seed cluster centers that have no feature vectors assigned to them (empty clusters) and re-index the cluster membership assignments accordingly to obtain K clusters.

Step 4: For each of the K clusters, compute the modified weighted fuzzy expected value [174] of all feature vectors in that cluster to be the new cluster center ($c(^k)$ for k = 1,...,K.

Step 5: If first pass, then go to Step 2;
else-if any clusters have changed, then go to Step 2;
else exit this part.

Step 6: Zero out all clust[q]; q = 1,...,Q

Step 7: Zero out all count[k]; k = 1,...,K

Step 8: For each q of the Q feature vectors do Find the nearest $c(^k)$ to q
Put clust[q] = k
Put count[k] = count[k] + 1

Step 9: If any cluster k^0 is empty, eliminate it by eliminating its cluster center $c(k^0)$and re-indexing the remaining cluster centers $c(^k)$ and the cluster counts (use clust[q] = k - 1 for all k where k ¿ k^0). Repeat this for each empty cluster.

Step 10: For each of the K clusters, compute the modified weighted fuzzy expected value vector of all feature vectors in that cluster.

Step 11: If any cluster has changed on this iteration, then go to Step 6.

Step 12: Save this clustering and exit this part.

Step 13: Find the two clusters that have the minimum distance between their cluster centers, replace the cluster center with the average of the two centers, re-index the cluster centers and reduce K accordingly.

Step 14: Do Steps 6 through 10 above to obtain a new clustering with one less cluster. for the new K clusters.

Step 15: Compute the Xie-Beni clustering validity value for the new K clusters.

Step 16: If the new Xie-Beni value is lower than the previous Xie-Beni value and K $>$ 2 then go to Step 13
else keep the previous clustering and stop.

The flowchart of IFC is given in Figure 3.12

3.5.3 Iris Data Set

Iris data set is a famous data set used for testing clustering and other learning algorithms. It is known to be noisy and inseparable [145]. It contains 150 feature vectors. Each feature vector is 4-dimensional representing the 4 iris features: petal width, petal length, sepal width, and sepal length. These feature vectors were labeled into 3 clusters, or species, which are Sestosa, Versicolor and Virginicus.

Figure 3.13 shows the petal and sepal of Iris flower. Figure 3.14 shows the three species of Iris flower. Figure 3.15 displays the spectramap biplot of Fisher's iris data set.

Figure 3.16 displays the Modified Weighted Fuzzy Expected Values (MWFEV) of each feature for each cluster. It shows that there is much overlapping of the second

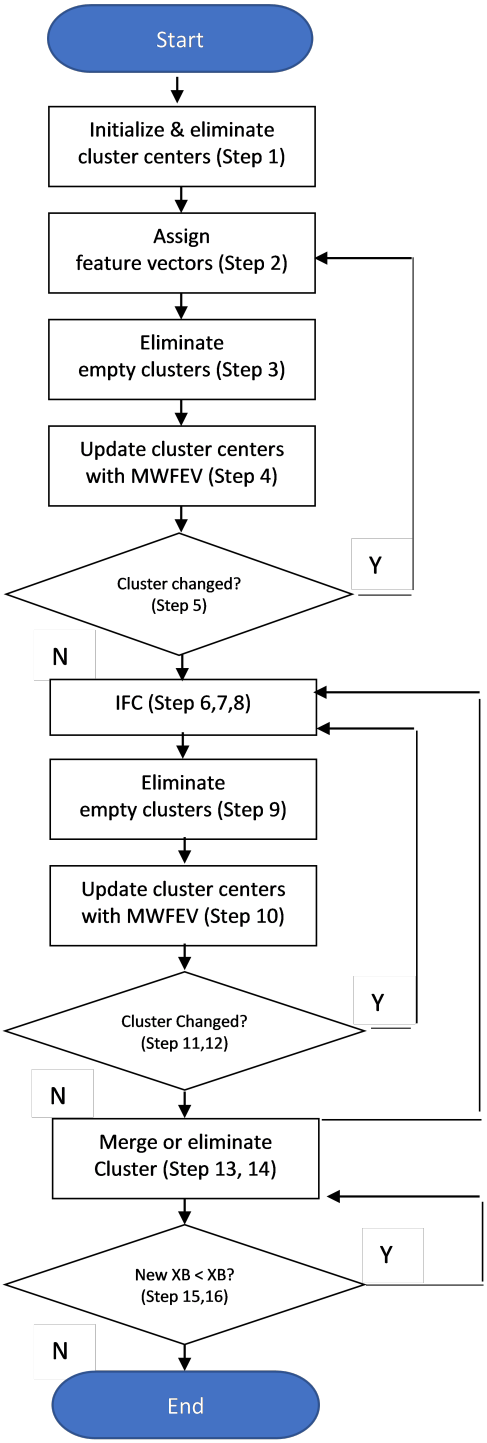

Figure 3.12: Flowchart of IFC

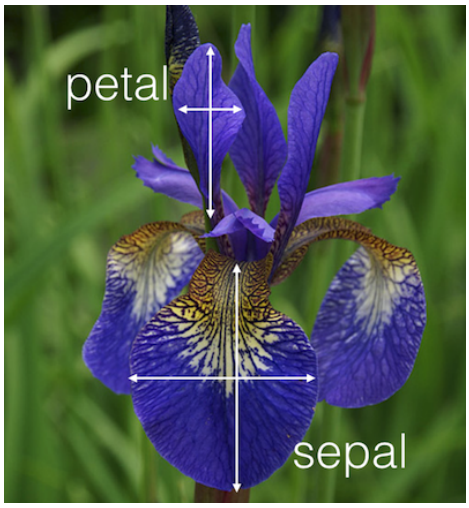

Figure 3.13: Petal and sepal of Iris flower [268]

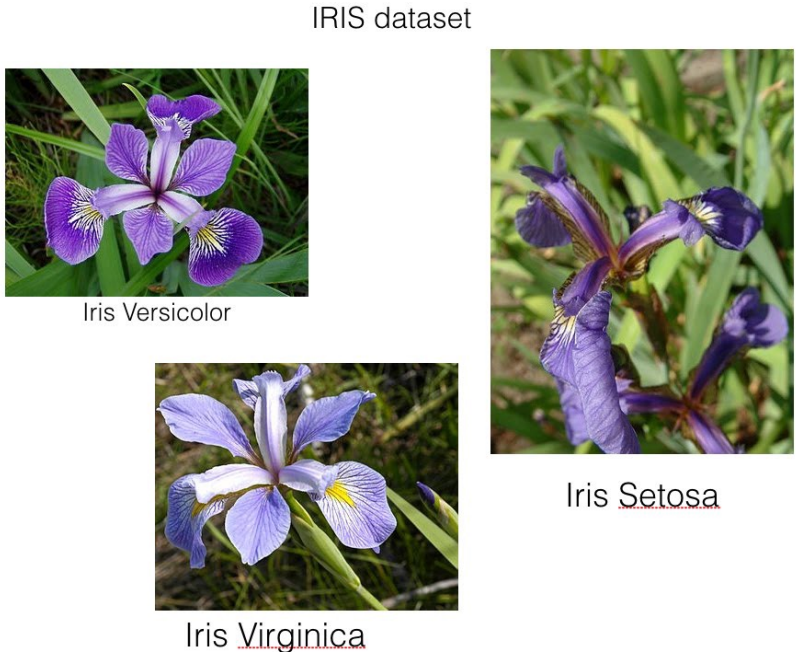

Figure 3.14: Three species of Iris flower [213]

feature values that do not differentiate the 3 clusters well. It shows that the third feature and the forth feature are the best separators, and the first feature is good separator [174].

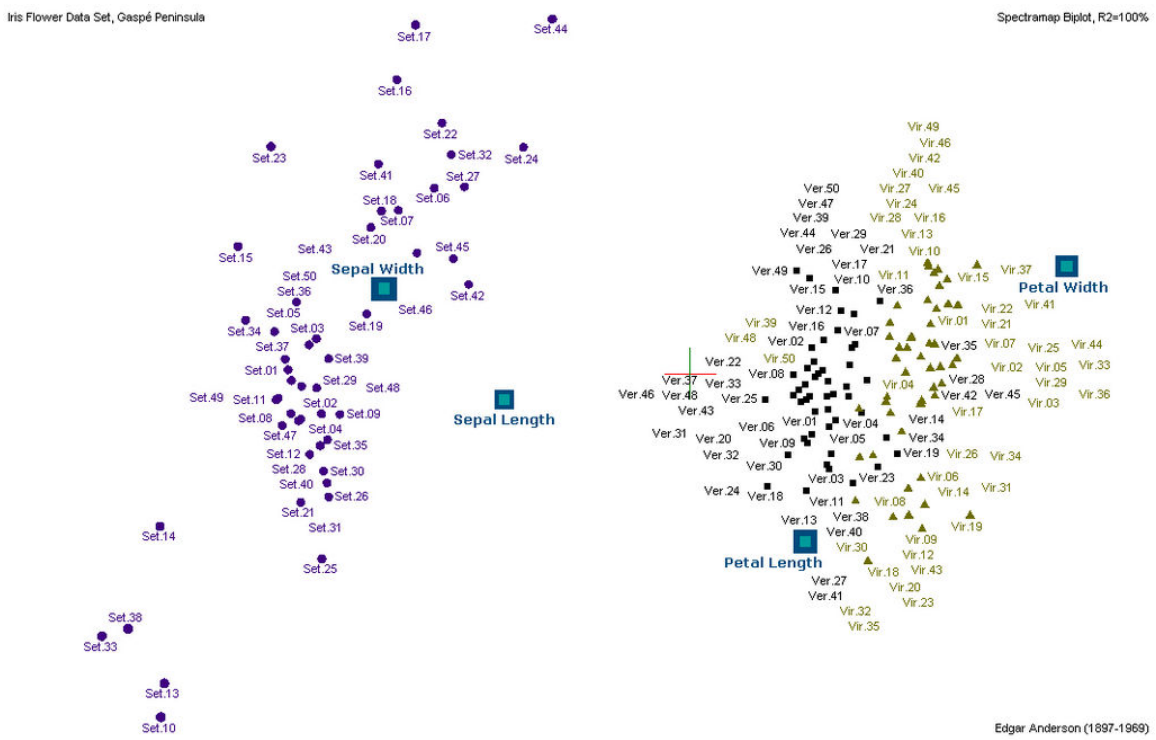

Figure 3.15: Spectramap biplot of Fisher's Iris data set [192]

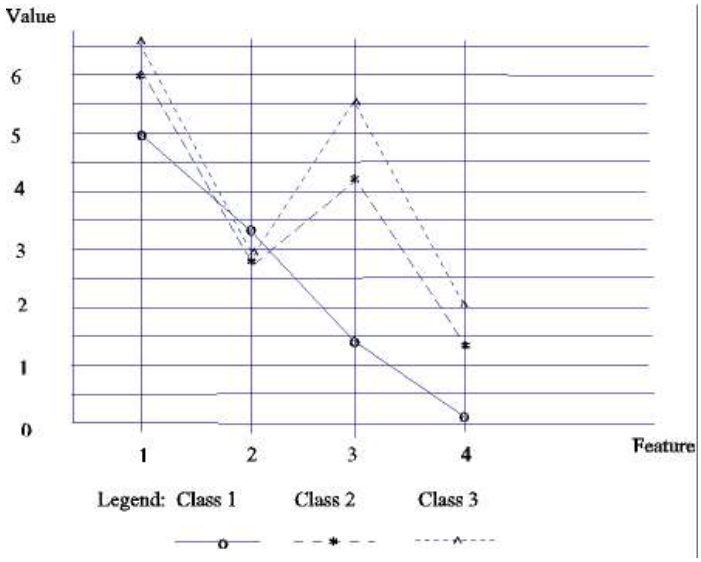

Figure 3.16: MWFEV centers of the four Iris features [174]

3.6 Improved Fuzzy Cluster Ensemble Algorithm

3.6.1 Diagram

Our proposed IFCE methodology is illustrated in Figure 3.17 (adapted from [264]). It includes two major steps: ensemble generation step generating base clusterings to

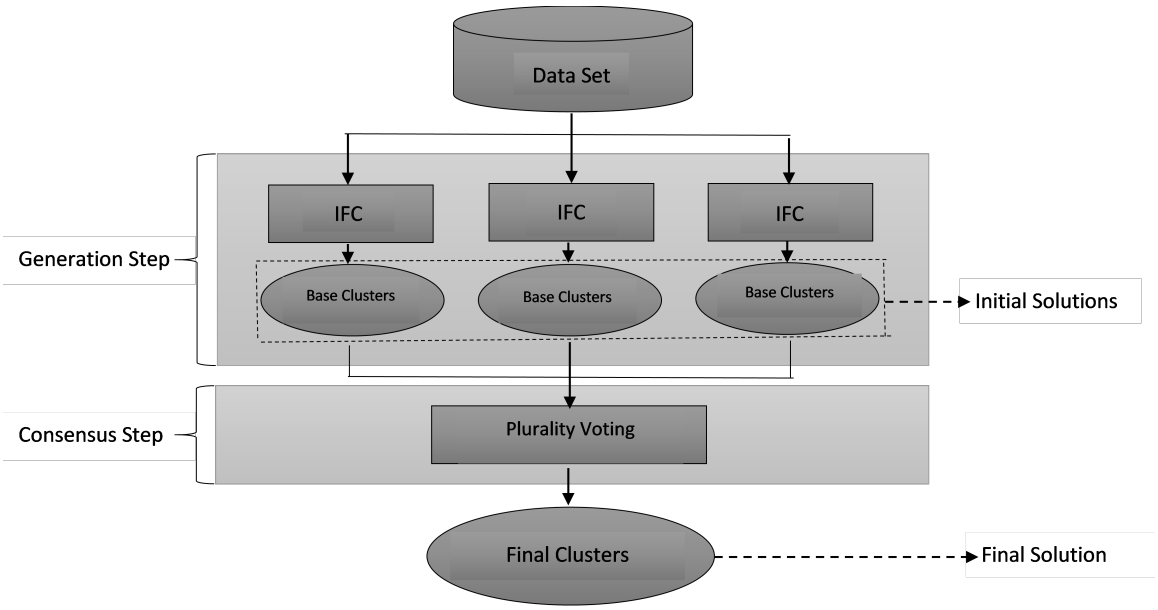

Figure 3.17: IFCE (adapted from [264])

form a cluster ensemble; and ensemble consensus step producing the final clustering result using a consensus function.

3.6.2 Ensemble Generation

In ensemble generation step, diversity is often artificially introduced in order to improve the output results of an ensemble. In homogeneous ensembles, based clusterings are created using a single clustering algorithm. In heterogeneous ensembles, base clusterings are created using different clustering algorithms. IFCE is a homogeneous ensemble, and its three base clusterings are Improved Fuzzy Clustering (IFC) using Modified Weighted Fuzzy Expected Value (MWFEV) [174] with different initializations.

With IFC, first a large number (twice the total existing feature vectors) seed cluster centers are generated in the feature space. Then those that are too close to obtain a reduced but uniformly distributed set of initial seeds are eliminated. For elimination, IFC uses the average distance between centers as a threshold, so that half of the seeds are eliminated. After IFC obtains the initial cluster centers, it applies

K-Means clustering (together, it is called Improved K-Means [173]).

After applying the Improved K-Means clustering, we use the MWFEV method for computing the cluster centers.The MWFEV is given by the equation [174]:

$$\vec{\mu}^{(r+1)} = \sum_{p=1}^{P} \alpha_p^{(r)} x_p \tag{3.13}$$

Where,

$$\alpha_p^{(r)} = \frac{exp[-\frac{(x_(p)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]}{\sum_{m=1}^{P} exp[-\frac{(x_(m)-\vec{\mu}^{(r)})^2}{(2\delta^2)^{(r)}}]} \tag{3.14}$$

$$(\delta^2)^{(r+1)} = \sum_{p=1}^{P} \alpha_p^{(r)} (x_p - \vec{\mu}^{(r)})^2 \tag{3.15}$$

After a number of iterations, IFC converges to many relatively small clusters ready for merging. In order to produce more natural shapes clusters as the results instead of forcing them into normed balls due to use of the distance function, IFC merges the closest clusters until the Xie-Beni validity measure does not decrease anymore or until the number of clusters is reduced to two. It finds the two clusters with the minimum distance between their centers, calculates a new center with the average of the two centers. It then reduces the number of clusters by one accordingly [174].

3.6.3 Ensemble Consensus

After diverse clustering results have been produced by the multiple base clustering algorithms, they need to be integrated into a single result. Voting method is commonly used as the consensus function in integrating clustering results for ensemble. IFCE uses plurality voting as its integration function to obtain the final clustering using clustering results from its three base clusterings. With plurality voting, each feature vector votes for or is assigned to one cluster in each base clustering, and the cluster who has more votes (plurality) than any other cluster is the winner. It is different than majority voting, with which the winner polls more than half of the votes.

Chapter 4

Experimental Results

In this chapter, we first describe our experimental design and settings including data sets, clustering validity measure, and other experimental conditions. We then present the experimental results on clustering criteria comparison analysis, parameters analysis, complexity analysis, and noise robustness analysis.

4.1 Experiment Design and Settings

The goal of our experiments is to compare the performance of IFCE to a number of clustering algorithms. We measure the performance in terms of clustering criteria, parameter sensitivities, complexity, and noise robustness. We choose eight real cancer gene expression data sets and ten synthetic noisy data sets for our comparison experiments.

4.1.1 Cancer Gene Expression Data Sets

We chose eight cancer gene expression data sets to evaluate our proposed cluster ensemble method, and they are described below.

Golub1999v1 data set contains the expression of 1,877 genes in 72 leukemia samples, with labels for 2 subtypes [104]. Golub1999v2 data set contains the expression of 1,877 genes in 72 leukemia samples, with labels for 3 subtypes [104]. These two Golub data sets are the most studied and cited microarray data set. Each of them contains 47 acute lymphoblastic leukemia (ALL) patients and 25 acute myeloid leukemia (AML)

patients. All the 72 patients had their bone marrow samples obtained at the time of diagnosis. The samples were assayed with Affymetrix Hgu6800 chips, and 7129 gene expressions (Affymetrix probes) were observed. Since ALL arises from two different types of lymphocytes (T-cell and B-cell), Leukemia1 can be considered containing three subtypes: AML, ALL-T, and ALL-B, as labeled in Leukemia2. [102].

Armstrong2002 data set contains the expression of 2,194 genes in 72 leukemia samples, with labels for 3 subtypes [14]. Initially, samples of 20 patients with conventional ALL (ALL) but without MLL translocation were collected. Then, samples from 17 patients with the MLL translocation (MLL) were collected. These samples were obtained from the peripheral blood or bone marrow of the patients at diagnosis or relapse. All these ALL and MLL patients were diagnosed as CD19+ B-precursor ALL by pathologists at the institution where the samples were collected [102].

Chowdary2006 BCT (Breast-Colon tumors) data set contains the expression of 182 genes in 104 samples, with labels for 2 subtypes [48]. It contains pairs of snap-frozen and RNAlater preservative-suspended samples from 30 such paired lymph node-negative breast cancer patients and 21 such paired Dukes' B colon cancer patients, as well as triplication of six stage II colon cancer patient samples.

Nutt2003 Brain Tumor data set contains the expression of 1,377 genes in 50 samples, with labels for 4 subtypes [210]. Samples were primary rumors and collected before therapy. They were reviewed by board-certified neuropathologists at the collecting hospital as 50 high-grade glioma samples: 28 glioblastomas and 22 anaplastic oligodendrogliomas. They were also reviewed by two additional neurologists for diagnostic confirmation. Classic glioblastomas were characterized by having irregularly distributed, pleomorphic, and hyperchromatic nuclei and sometimes with conspicuous eosinophilic cytoplasm. Anaplastic oligodendrogliomas were designated as having classic histopathology exhibiting relatively evenly distributed, uniform, and rounded nuclei and frequent perinuclear halos.

Pomeroy2002 CNS (central nervous system) data set contains the expression of 1,379 genes in 42 samples, with labels for 5 subtypes [222]. Patients include 10

medulloblastomas, 10 malignant gliomas (WHO grades III and IV), 5 AT/RT, 5 renal/extrarenal rhabdoid tumors, 8 supratentorial PNETs, and 4 normal cerebella). All samples were obtained at initial surgery prior to treatment. Affymetrix scanners were used to scan the arrays and the gene expression values were calculated by Affymetrix GENECHIP software. A variation filter was applied to exclude genes showing minimal variation across the samples. The data were also normalized by standardizing each sample to mean 0 and variance 1.

Chen2002 HCC (hepatocellular carcinoma) data set contains the expression of 85 genes in 180 samples, with labels for 2 subtypes [44]. It includes 102 primary HCC (from 82 patients), 74 nontumor liver tissues (from 72 patients), seven benign liver tumor samples (three adenoma and four FNH), 10 metastatic cancers, and 10 HCC cell lines. Each sample was independently reviewed by two pathologists. The array was scanned using GenePix 4000A microarray scanner by Axon Instruments.

Khan2001 SRBCT (small, round blue-cell tumors) data set contains the expression of 1,069 genes in 83 samples, with labels for 4 subtypes [147]. It contains gene expression profiles of four types of childhood SRBCT: neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), a subset of non-Hodgkin lymphoma, and the Ewing family of tumors (EWS). All the histological diagnoses were made at hospitals extensively experienced in diagnosing pediatric cancers. Standard NHGRI protocol was followed to obtain expression images, which were then analyzed by DeArray software39. Each sample was normalized across all experiments. The natural logarithm (ln) of was applied to obtain the value of the expression levels.

The eight real cancer gene expression data sets used for experiments are summarized in Table 4.1. They are filtered data sets from the empirical study of de Souto et al. [64] with uninformative genes removed. They were originally obtained from published microarray studies.

Data Set	Cancer Type	Samples	Genes	Clusters	Chip
Golub1999v1	Leukemia (bone marrow)	72	1,877	2	Affy.
Golub1999v2	Leukemia (bone marrow)	72	1,877	3	Affy.
Amstrong2002	Leukemia (bone marrow)	72	2,194	3	Affy.
Chowdary2006	Breast-Colon Tumors (breast and colon)	104	182	2	Affy.
Nutt2003	Brain Tumor (brain)	50	1,377	4	Affy.
Pomeroy2002	Central Nervous System (brain)	42	1,379	5	Affy.
Chen2002	Hepatocellular Carcinoma (liver)	180	85	2	cDNA
Khan2001	Small, Round Blue-cell Tumors (multi-tissue)	83	1,069	4	cDNA

Table 4.1: Cancer gene expression data sets

4.1.2 Comparable Clustering Algorithms

The algorithms used for comparison are as follows: a) four traditional or simple clustering algorithms: KM (K-Means), SL (Single-Linkage), CL (Complete-Linkage), AL (Average-Linkage), and b) six state-of-the-art cluster ensemble methods: MULTI-K, CCHC (Consensus Clustering with Hierarchical Clustering), GCC (Graph-Based Consensus Clustering), CSPA (Cluster-Based Similarity Partitioning Algorithm), HGPA (Hyper-Graph Partitioning Algorithm), MCLA (Meta-Clustering Algorithm).

MULTI-K and CCHC was designed for analyzing low-sample and high-dimensional gene expression data with high level of accuracy compared with k-means or hierarchical clusterings [148]. GCC was the first time in which graph-based cluster ensemble is applied to cluster discovery for microarray data with better performance than most existing algorithms [295]. CSPA, HGPA and MCLA are well-known graph-based cluster ensemble benchmarks in the literature [279].

Kim et al. developed MULTI-K, which combines multiple K-Means runs with varied number of clusters. It uses the single link agglomerative hierarchical clusterings as the consensus function [148]. Monti et al. proposed CCHC (consensus clustering with hierarchical clustering) method that uses the average link agglomerative hierarchical clusterings as the consensus function [197]. Yu et al. designed graph based consensus clustering (GCC) that repeats subspace generation and subspace clustering to obtain different clusterings for calculating the final consensus clustering [295]. Strehl and Ghosh proposed three methods that use a graph partitioning algorithm

called METIS [140] to partition a similarity graph (or hyper-graph, meta-level graph) to obtain final consensus clustering [259]. Cluster-based Similarity Partitioning Algorithm (CSPA) constructs a similarity graph where vertices represent samples and edge weights represent similarity based on a co-association matrix. HyperGraph Partitioning Algorithm (HGPA) constructs a hyper-graph where vertices represent samples and the hyper-edges (same-weighted) represent clusters in the ensemble. Meta-Clustering Algorithm (MCLA) constructs a meta-level graph where vertices represent clusters in the ensemble and edge weights represent binary Jaccard measures [127].

Four comparable clustering algorithms used in experiments are presented in Table 4.2.

Six comparable cluster ensemble algorithms used in experiments are presented in Table 4.3.

Algorithm	Acronym&Reference	Category	Time Complexity
K-means	KM [258]	Partitioning clustering	$O(N)$
Single-linkage	SL [105]	Hierarchical clustering	$O(N^2)$
Complete-linkage	CL [66]	Hierarchical clustering	$O(N^2 \log N)$
Average-linkage	AL [61]	Hierarchical clustering	$O(N^2 \log N)$

Table 4.2: Comparable clustering algorithms

Algorithm	Acronym& Reference	Consensus Function	Time Complexity
Multi-K	Multi-K [148]	Pairwise Similarity	$O(KNM)$
Consensus Clustering with Hierarchical Clustering	CCHC [197]	Pairwise Similarity	$O(N^3M)$
Graph-based Consensus Clustering	GCC [295]	Pairwise Similarity	$O(KNM)$
Cluster-based Similarity Partitioning Algorithm	CSPA [259]	Graph-based	$O(N^2KM)$
HyperGraph Partitioning Algorithm	HGPA [259]	Graph-based	$O(NKM)$
Meta-Clustering Algorithm	MCLA [259]	Graph-based	$O(NK^2M^2)$

Table 4.3: Comparable cluster ensemble algorithms

4.1.3 Validity Measure

Evaluating the quality of a clustering result is difficult and ill-posed. Unlike supervised learning, clustering tasks usually do not have cluster labels of the data objects available. Thus, determining which clustering result is better becomes difficult.

Clustering validity measure shows how well the clustering performs in detecting the underlying patterns of the data objects, possibly with respect to the hidden true labels on these data objects when they are available. A number of objective measures can be used to quantify the quality of the clusters obtained by different clustering methods [286, 113]. Intrinsic validity measures evaluate the result based on information intrinsic to the data alone. External validity measures evaluate the result based on previous knowledge about the data. Some common external validity measures are shown in Figure 4.1. Some common internal validity measures are shown in Figure 4.2. Some common relative validity measures are shown in Figure 4.3.

Evaluation Index	Implication	Formula	Notes
Purity	Reflects the purity of objects in clusters	$\sum_{i=1}^{k}\frac{1}{N}\max_{j}(n_i^j)$	n_i^j is the number of objects in the *i*th cluster belonging to the *j*th category
Entropy	Reflects the confounding, or impurity, of objects in clusters	$\sum_{i=1}^{k}\frac{n_i}{N}(-\frac{1}{\log l}\sum_{j=1}^{l}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i})$	See Purity
Rand	Measures the similarity between clustering result C and the a priori partition P	$(a+d)/(a+b+c+d)$	a, b, c and d indicate the number of object pairs in SS, SD, DS and DD, respectively,and $a+b+c+d=N(N-1)/2$
Jaccard	Measures the similarity between C and P	$a/(a+b+c)$	See Rand
FM	Measures the similarity between C and P	$\sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}}$	See Rand

D indicates the dataset containing *N* data objects, and its a priori partition with *l* categories is $P=\{P_1,P_2,\ldots,P_l\}$. The result obtained from clustering is $C=\{C_1,C_2,\ldots,C_k\}$, where *k* is the number of clusters. SS indicates that both objects belong to the same cluster of *C* and to the same group of partitions *P*; SD indicates points that belong to the same cluster of *C* and to different groups of *P*; DS indicates points that belong to different clusters of *C* and to the same group of *P*; and DD indicates points that belong to different clusters of *C* and to different groups of *P*.
doi:10.1371/journal.pone.0090109.t001

Figure 4.1: Common external validity measures [242]

Evaluation Index	Implication	Formula	Notes
CPCC	Measures the similarity between matrix Pc and P, where Pc(i,j) is the similarity of x_i and x_j when they are assigned to the same cluster, and P is an adjacency matrix	$\frac{1/M\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij}c_{ij}-\mu_P\mu_C)}{\sqrt{\left[1/M\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij}^2-\mu_P^2)\right]\left[1/M\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_{ij}^2-\mu_C^2)\right]}}$, where *N* is the number of objects, and $M=N\cdot(N-1)/2$, $\mu_p=(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}P(i,j)$,$\mu_C=(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}P_C(i,j)$, d_{ij},c_{ij} represent the elements in P and Pc, respectively.	$-1\le CPCC\le 1$. This is appropriate for hierarchical clustering.
Hubert's Γ	Measures the similarity between a clustering result C and the proximity matrix P	$(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}P(i,j)C(i,j)$, where $C(i,j)=\begin{cases}1, if\ c(x_i)\neq c(x_j)\\ 0, otherwise\end{cases}$	The larger the value, the higher the similarity between P and C.
Normalized Γ	Measures the similarity between a clustering result C and the proximity matrix P	$\frac{(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(P(i,j)-\mu_P)(C(i,j)-\mu_C)}{\sigma_P\sigma_C}$	The larger the value, the higher the similarity between P and C.

doi:10.1371/journal.pone.0090109.t002

Figure 4.2: Common internal validity measures [242]

To evaluate the clustering results of IFCE against the other clustering algorithms, external clustering validity measure of Classification Accuracy (CA) [206] is chosen

Evaluation Index	Implication	Formula	Notes
Dunn	Measures the compactness of clusters and separation between clusters	$\min_{i=1,\ldots,n_c}\left\{\min_{j=i+1,\ldots,n_c}\left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,n_c} diam(c_k)}\right)\right\}$	The larger the value, the better the clustering effect
DB	Measures the compactness of clusters and separation between clusters	$DB=\frac{1}{n_c}\sum_{i=1}^{n_c} R_i,\ (R_i=\max_{i=1,\ldots,n_c,i\neq j} R_{ij}, i=1,\ldots,n_c;\ R_{ij}=(s_i+s_j)/d_{ij})$	The smaller the value, the better the clustering effect
RMSSDT	Measures the differences between clusters	$\frac{\sum_{i=1,\ldots,n_c}\sum_{j=1,\ldots,d}\sum_{k=1}^{n_{ij}}(x_k-\bar{x}_k)^2}{\sum_{i=1,\ldots,n_c}\sum_{j=1,\ldots,d}(n_{ij}-1)}$	The smaller the value, the better the clustering effect
SD	Measures the compactness of clusters and separation between clusters	$SD(n_c)=\alpha\cdot scatt(n_c)+dis(n_c),\ scatt(n_c)=\frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\|\sigma(v_i)\|}{\|\sigma(X)\|},\ dis(n_c)=\frac{d_{\max}}{d_{\min}}\sum_{k=1}^{n_c}\left(\sum_{z=1}^{n_c}\|v_k-v_z\|\right)^{-1}$	The larger the value, the better the clustering effect
S_Dbw	Measures the intra-cluster variance and inter-cluster density	$S\_Dbw(n_c)=scat(n_c)+dbw(n_c),\ dbw(n_c)=\frac{1}{n_c(n_c-1)}\sum_{i=1}^{n_c}\sum_{\substack{j=1\\ i\neq j}}^{n_c}\frac{d(u_{ij})}{\max\{d(v_i),d(v_j)\}}$	The smaller the value, the better the clustering effect

doi:10.1371/journal.pone.0090109.t003

Figure 4.3: Common relative validity measures [242]

because of the importance of domain meaningfulness. CA calculates the percentage of accurately clustered data objects among all data objects clustered. Let Q be the number of total clustered objects, and a be the number of accurately clustered objects. CA is defined by the equation:

$$CA = \frac{a}{Q} * 100\% \tag{4.1}$$

Higher value of CA means higher clustering accuracy.

CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM from the study of Iam-on et al. (supplementary data) [124] are adopted for evaluating against CA results of IFCE.

Cluster labels are available in the data sets, but they are not used in any clustering process in the experiments. Cluster labels are only used to calculate CA after clustering is finished.

We run IFCE over the data sets. The outputs include cluster assignment for each data object and the CA value for each run.

Except for the experiments of parameter analysis on ensemble size M, each clustering method repeats for 50 runs and the average of CA values is adopted. This approach helps to reduce the effect of stochastic variation with clustering methods

Data Set	IFCE	MULI-K	CCHC	GCC	CSPA	HGPA	MCLA	SL	CL	AL	KM
Golub1999v1	0.743	0.656	0.726	0.738	0.686	0.716	0.718	0.667	0.653	0.653	0.731
Golub1999v2	0.625	0.544	0.673	0.693	0.612	0.710	0.685	0.542	0.542	0.542	0.669
Armstrong2002	0.627	0.607	0.780	0.820	0.770	0.775	0.821	0.403	0.472	0.403	0.761
Chowdary2006	0.923	0.654	0.654	0.683	0.860	0.875	0.898	0.606	0.606	0.606	0.654
Nutt2003v2	0.557	0.602	0.619	0.619	0.622	0.597	0.644	0.360	0.480	0.360	0.613
Pomeroy2002	0.403	0.536	0.595	0.640	0.581	0.579	0.540	0.333	0.405	0.333	0.581
Chen2002	0.799	0.581	0.587	0.608	0.836	0.834	0.821	0.581	0.581	0.581	0.599
Khan2001	0.543	0.457	0.433	0.434	0.443	0.462	0.457	0.373	0.349	0.361	0.443

Table 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].

and achieve consistency in clustering results.

4.1.4 Number of Clusters

For each of the six cluster ensembles chosen to evaluate against IFCE, fixed number of clusters (K) with full space data is used. For IFCE, automatic calculated K with full space data is used. IFCE uses Improved K-Means method that can find K automatically by reducing a large initial K via merging small clusters. IFCE no longer needs a user specified fixed K to perform clustering as many other clustering methods require.

4.2 Validity Measure Comparison

The CA results of IFCE and other investigated clustering algorithms on real cancer gene expression data sets are presented in Table 4.4 and Figure 4.4 through Figure 4.12.

As we mentioned earlier, CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM from the study of Iam-on et al. (supplementary data) [124] are adopted for evaluating against CA results of IFCE.

Figure 4.4 illustrates that IFCE is the top performer on three of the eight data sets, more than any other methods examined. Also, it performs well on most of the

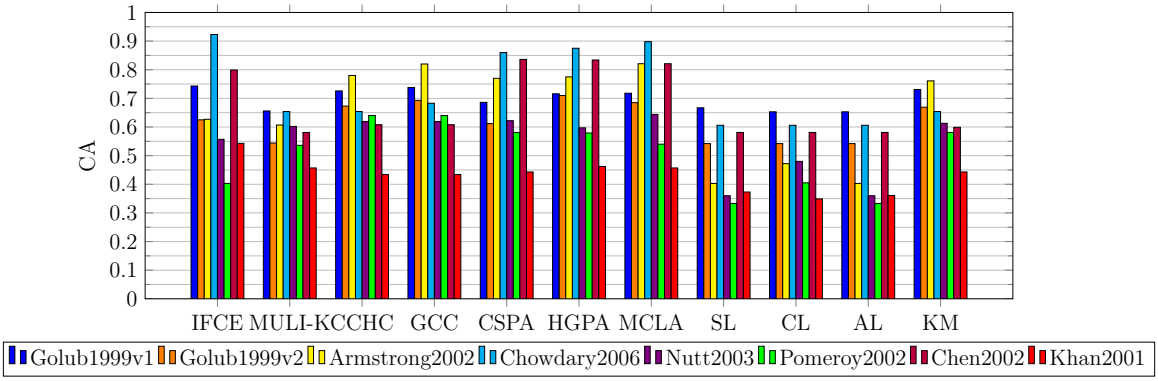

Figure 4.4: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM across eight real cancer gene expression data sets over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

other data sets.

Figure 4.5 illustrates that IFCE is the top 1 performer with CA value of 0.743 for data set Golub1999v1 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM.

Figure 4.6 illustrates that IFCE is the 6th performer with CA value of 0.625 for data set Golub1999v2 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM. The top 1 performer is HGPA with CA value of 0.710.

Figure 4.7 illustrates that IFCE is the 7th performer with CA value of 0.627 for data set Armstrong2002 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM. The top 1 performer is MCLA with CA value of 0.821.

Figure 4.8 illustrates that IFCE is the top 1 performer with CA value of 0.923 for data set Chowdary2006 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM.

Figure 4.9 illustrates that IFCE is the 8th performer with CA value of 0.557 for data set Nutt2003 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM. The top 1 performer is MCLA with CA value of 0.644.

Figure 4.10 illustrates that IFCE is the 9th performer with CA value of 0.403 for data set Pomeroy2002 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL,

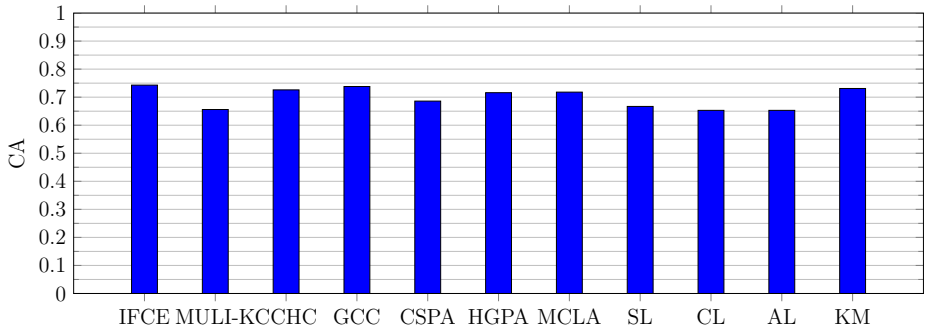

Figure 4.5: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Golub1999v1 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

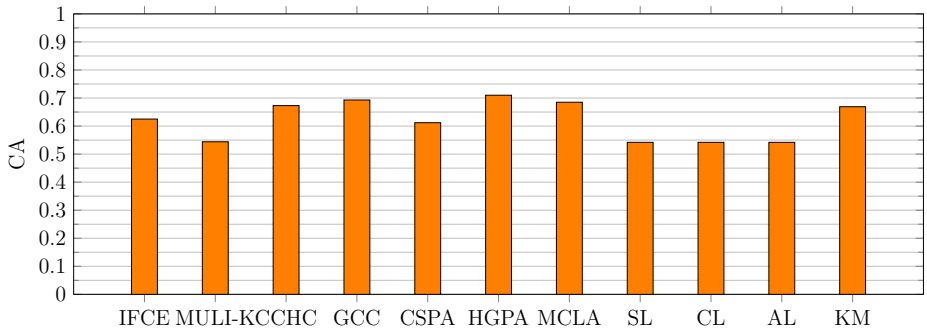

Figure 4.6: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Golub1999v2 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

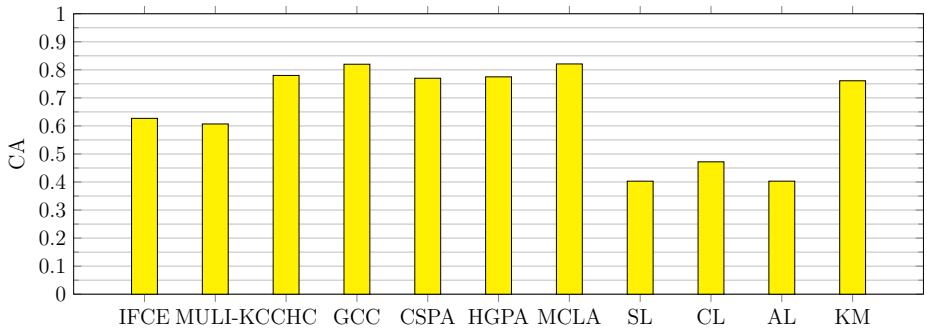

Figure 4.7: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Armstrong2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

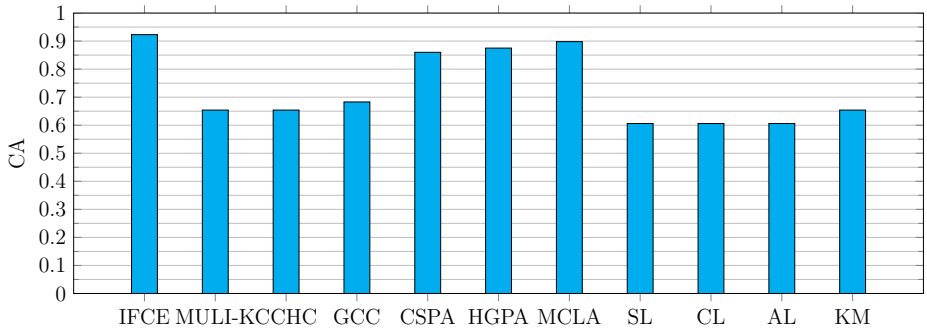

Figure 4.8: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Chowdary2006 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

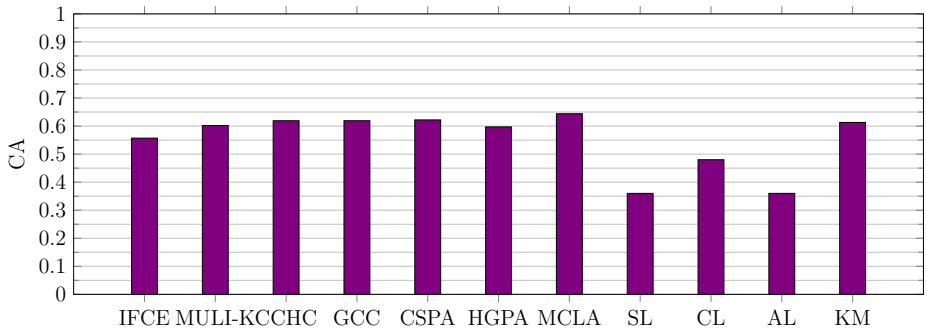

Figure 4.9: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Nutt2003 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

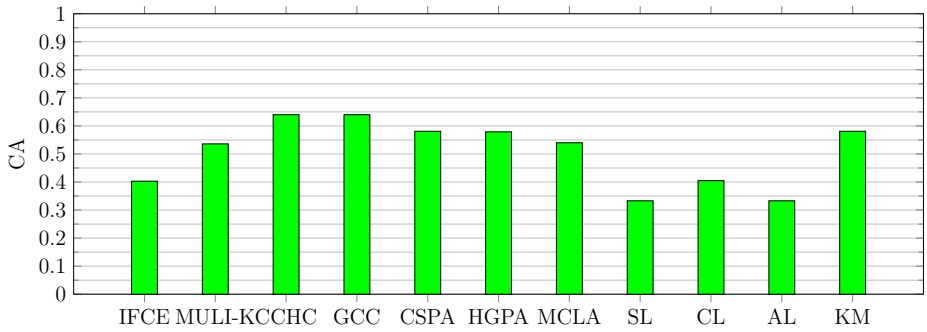

Figure 4.10: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Pomeroy2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

CL, AL, and KM. The top 1 performers are CCHC and GCC both with CA values of 0.640.

Figure 4.11 illustrates that IFCE is the 4th performer with CA value of 0.799 for data set Chen2002 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM. The top 1 performer is CSPA with CA value of 0.836.

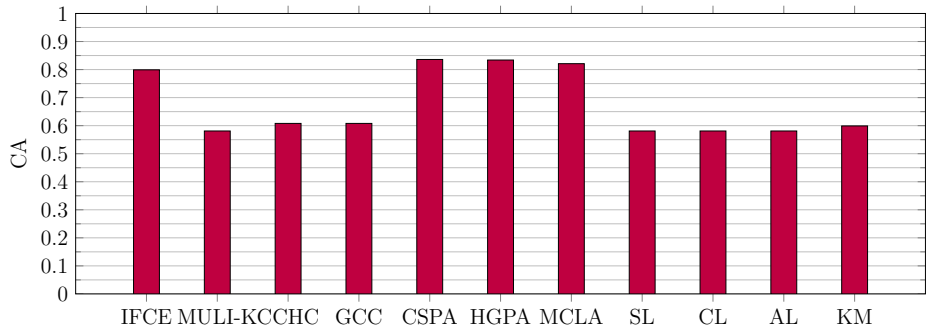

Figure 4.11: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Chen2002 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

Figure 4.12 illustrates that IFCE is the top 1 performer with CA value of 0.543 for data set Khan2001 among MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM.

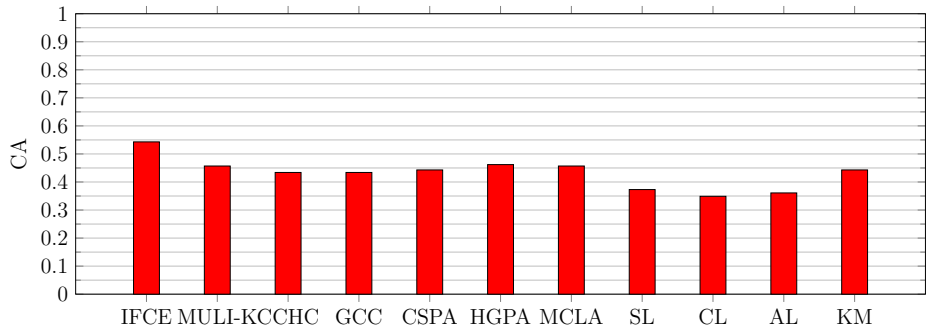

Figure 4.12: CA (Classification Accuracy) of IFCE, MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, KM for data set Khan2001 over 50 runs of each algorithm. CA results of MULTI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM are adopted from the study of Iam-on et al. (supplementary data) [124].

4.3 Parameter Analysis

IFCE provides the option of defining the values of multiple parameters. The initial values we define in our experiments are based on empirical experience or general estimation, and they work well on our data sets. However, in an explorative study, it is helpful to experiment with various values. This way, we can assess if IFCE has high degree of dependency on any particular values of some parameters. We can also find out the relation between IFCE and its parameters.

To evaluate IFCE's performance on various parameter values, parameter analysis are examined next. We choose two of the eight data sets for parameter analysis due to paper space constraints. Based on Figure 4.4, we select data set Chowdary2006 because IFCE produces the highest CA value on it. In addition, we select data set Chen2002 because IFCE produces one of the relatively average CA values on it.

4.3.1 N (number of clustering runs)

The first parameter examined is the number of clustering runs N. Smaller number of runs saves computing time, however it may not be enough runs to achieve the desired accuracy due to stochastic variation. Larger number of runs have the potential to increase clustering accuracy, but it increases the expense in run time. Therefore, various values of N are chosen. CA results with N = 1, 5, 50, 100, 200 for data sets Chowdary2006 and Chen2002 are presented in Table 4.5 and Figure 4.13. Run Time (Sec.) results are presented in Table 4.6 and Figure 4.14.

Data Set	N(1)	N(5)	N(50)	N(100)	N(200)
Chowdary2006	0.913	0.923	0.923	0.924	0.930
Chen2002	0.782	0.798	0.799	0.797	0.797

Table 4.5: CA of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.

Table 4.5, Figure 4.13, Table 4.6, and Figure 4.14 show that CA values of IFCE are relatively stable with varying N values. They also show that higher N values

Data Set	N(1)	N(5)	N(50)	N(100)	N(200)
Chowdary2006	58	282	2,846	5,672	11,326
Chen2002	134	688	6,991	13,087	26,296

Table 4.6: Run Time (sec.) of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.

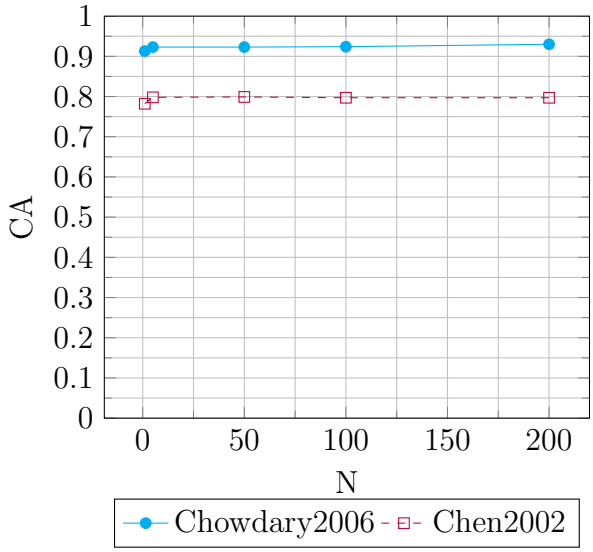

Figure 4.13: CA of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.

usually produce small increase in clustering accuracy while increase run time.

4.3.2 IMT (initial merging threshold)

The second parameter examined is the initial merging threshold (IMT) used to merge small clusters. Theoretically, when the IMT is small, clusters with shorter distance in between are merged while clusters with longer distance in between are not. Although the merging threshold increases during next clustering iteration to merge clusters with longer distance, the run time is longer than if we had chosen a larger initial merging threshold. However if we use a larger value as the initial merging threshold, we may risk missing small clusters by merging them at the beginning. Therefore, various values of IMT are chosen. CA results with IMT = 1.0, 2.0, 3.0, 4.0 for data sets Chowdary2006 and Chen2002 are presented in Table 4.7 and Figure 4.15. Run

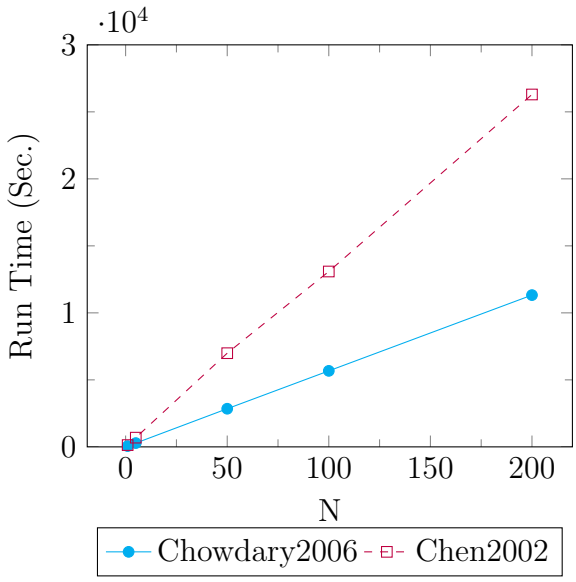

Figure 4.14: Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with N (number of clustering runs) = 1, 5, 50, 100, 200.

Time (Sec.) results are presented in Table 4.8 and Figure 4.16.

Data Set	IMT(1.0)	IMT(2.0)	IMT(3.0)	IMT(4.0)
Chowdary2006	0.923	0.921	0.918	0.926
Chen2002	0.799	0.796	0.800	0.800

Table 4.7: CA of IFCE on Chowdary2006 and Chen2002 with IMT(initial merging threshold) = 1.0, 2.0, 3.0, 4.0.

Data Set	IMT(1.0)	IMT(2.0)	IMT(3.0)	IMT(4.0)
Chowdary2006	2,846	2,820	2,879	2,848
Chen2002	6,991	6,633	6,613	6,519

Table 4.8: Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with IMT(initial merging threshold) = 1.0, 2.0, 3.0, 4.0.

Table 4.7, Figure 4.15, Table 4.8, and Figure 4.16 show that CA values of IFCE change within about 0.008 and run time values are relatively flat with varying IMT values.

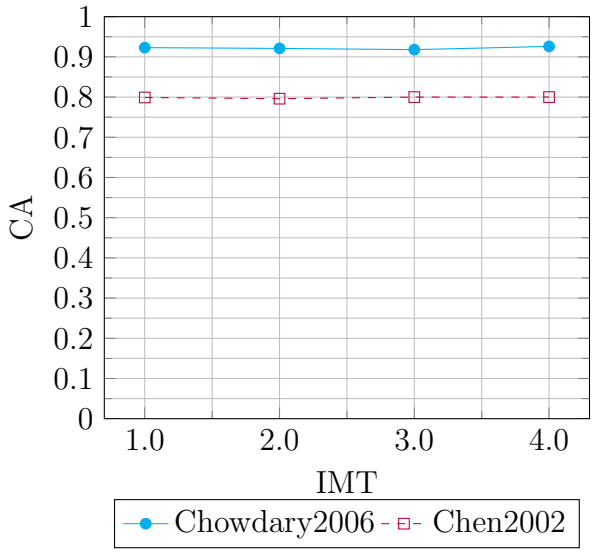

Figure 4.15: CA of IFCE on Chowdary2006 and Chen2002 with IMT (initial merging threshold) = 1.0, 2.0, 3.0, 4.0.

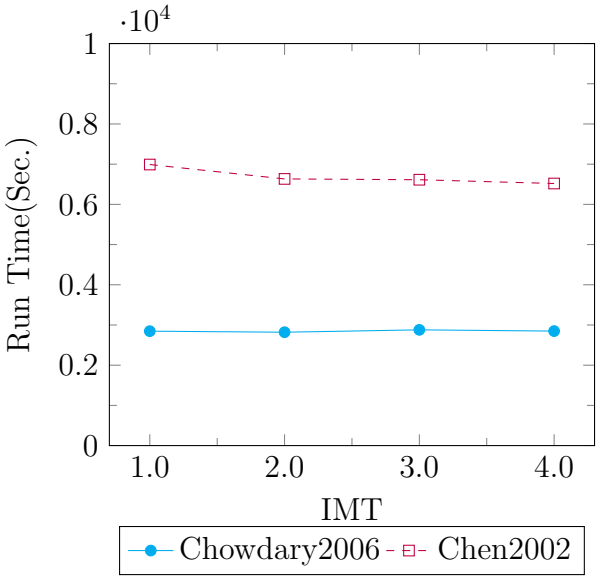

Figure 4.16: Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with IMT (initial merging threshold) = 1.0, 2.0, 3.0, 4.0.

4.3.3 M (ensemble size)

The third parameter examined is the ensemble size (M), or the number of base clusterings. From past studies [114], increasing ensemble size leads to increased clustering accuracy, although it is possible that the improvement becomes plateau when ensem-

ble size reaches certain value. Larger ensemble size also causes longer run time. A balance is desired between accuracy and run time. Therefore, various values of M are chosen. CA results with M = 3, 7, 11, 21 for data sets Chowdary2006 and Chen2002 are presented in Table 4.9 and Figure 4.17. Run Time (Sec.) results are presented in Table 4.10 and Figure 4.18.

Data Set	M(3)	M(7)	M(11)	M(21)
Chowdary2006	0.923	0.932	0.942	0.942
Chen2002	0.799	0.793	0.800	0.816

Table 4.9: CA of IFCE on Chowdary2006 and Chen2002 with M(ensemble size) = 3, 7, 11, 21.

Data Set	M(3)	M(7)	M(11)	M(21)
Chowdary2006	58	139	216	411
Chen2002	134	314	494	940

Table 4.10: Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with M(ensemble size) = 3, 7, 11, 21.

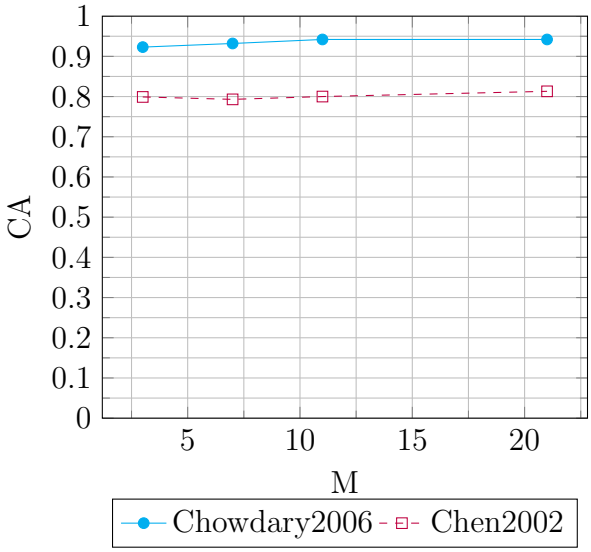

Figure 4.17: CA of IFCE on Chowdary2006 and Chen2002 with M (ensemble size) = 3, 7, 11, 21.

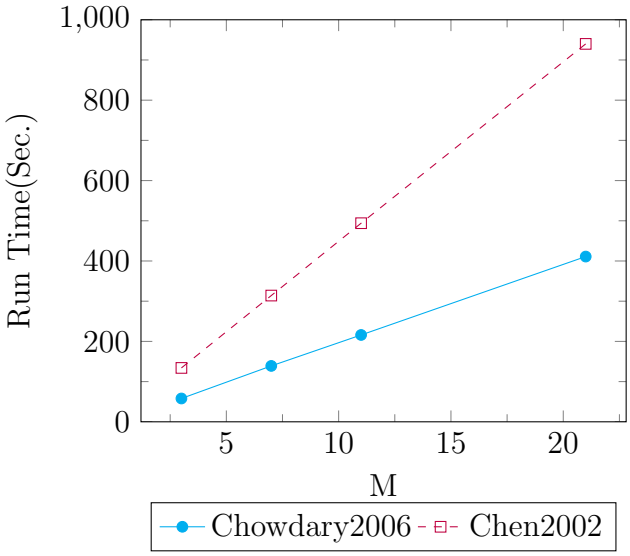

Figure 4.18: Run Time (Sec.) of IFCE on Chowdary2006 and Chen2002 with M (ensemble size) = 3, 7, 11, 21.

Table 4.9, Figure 4.17, Table 4.10, and Figure 4.18 show that CA values of IFCE are relatively stable with varying M values. They also show that higher M values produce small increase in clustering accuracy while increase run time.

4.4 Complexity Analysis

To evaluate IFCE's performance on various complexity conditions, time and space complexity analysis are examined next.

Let Q be the number of examples in the data set, N be the number of dimensions in the data set, K be the number of clusters, and M be the number of base clusterings.

4.4.1 Time Complexity

When we examine time complexity, IFCE involves three stages. **Stage 1:** initialization with I iterations of Improved K-Means. One vector distance costs $O(N)$, and complexity for KQ distances is $O(KQN)$. Complexity for I iterations is $O(IKQN)$. **Stage 2:** three base clusterings. One vector distance costs $O(N)$, and complexity for KQ distances is $O(KQN)$. Cost for computing the weights for K cluster centers is

$O(KQ)$. The complexity for M base clusterings is $O(MKQN)$. **Stage 3:** relabeling and plurality voting ensemble. Relabeling and voting approach is proved to be $O(K^3)$ [76]. The time complexity of IFCE is $O(IKQN)$+$O(MKQN)$+$O(K^3)$. So, IFCE's time complexity converges to $O(N)$.

4.4.2 Space Complexity

When we examine space complexity, for each base clustering, the cost of storing a matrix of QxN in memory is $O(QN)$. For M base clusterings, the total cost is $O(MQN)$. The cost of storing relabeling matrix is $O(K^2)$. The space complexity of IFCE is $O(MQN)$+$O(K^2)$, and converges to $O(N)$.

4.5 Noise Robustness Analysis

To examine the boundaries of IFCE's ability to maintain homogeneous clusters under conditions involving high noise-to-signal ratio data sets, we have created ten synthetic noisy data sets. The clustering process is repeated 50 times on each data set and the resulting clusterings at different noise levels were examined.

4.5.1 Synthetic Noisy Data Sets

The ten synthetic noisy data sets are based on real data sets Chowdary2006 and Chen2002 with increasing noise-to-signal ratios. Noise is incorporated by adding a constant (the maximum value in the gene) to the expression of cancer samples for that gene, such that the percentage of cancer samples with such added noise is 10%, 20%, 30%, 40%, 50%. Such cancer samples represent outliers in the data sets. The ten synthetic noisy data sets used for experiments are summarized in Table 4.11 and Table 4.12.

4.5.2 Results

The performance of IFCE on synthetic noisy data sets are presented in Table 4.13 and Figure 4.19.

Data Set	Noise%	Samples	Genes	Clusters
Chowdary2006_10p	10%	104	182	2
Chowdary2006_20p	20%	104	182	2
Chowdary2006_30p	30%	104	182	2
Chowdary2006_40p	40%	104	182	2
Chowdary2006_50p	50%	104	182	2

Table 4.11: Synthetic noisy data sets created by adding artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% to Chowdary2006

Data Set	Noise%	Samples	Genes	Clusters
Chen2002_10p	10%	180	85	2
Chen2002_20p	20%	180	85	2
Chen2002_30p	30%	180	85	2
Chen2002_40p	40%	180	85	2
Chen2002_50p	50%	180	85	2

Table 4.12: Synthetic noisy data sets created by adding artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% to Chen2002

Data Set	Noise					
	0%	10%	20%	30%	40%	50%
Chowdary2006	0.923	0.892	0.879	0.877	0.893	0.890
Chen2002	0.799	0.771	0.780	0.793	0.774	0.775

Table 4.13: Noise Robustness with artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% added to Chowdary2006 and Chen2002.

Table 4.13 and Figure 4.19 show that IFCE demonstrates robustness to highly noisy data sets. It maintains cluster classification accuracy above 0.870 for Chowdary2006 and above 0.770 for Chen2002 even when signals are reduced by 50%.

4.6 Conclusion

We have presented a new fuzzy cluster ensemble method IFCE. We have also evaluated IFCE through comparisons with numerous existing benchmark ensemble clustering and simple clustering methods on eight real cancer gene expression data sets. IFCE is the top performer on three of the eight data sets, more than any other methods

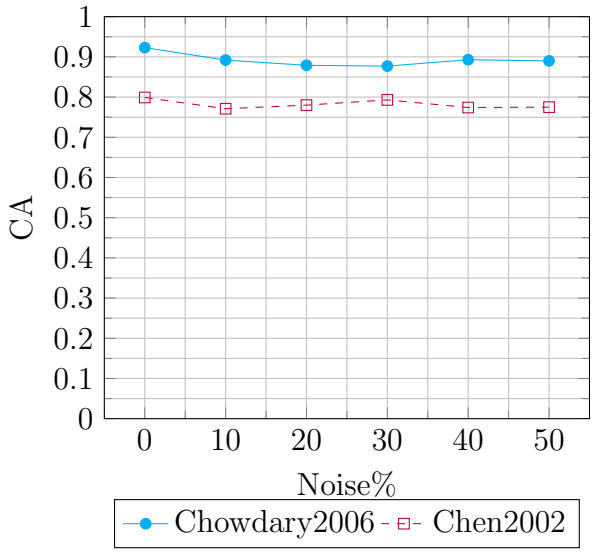

Figure 4.19: Noise Robustness with artificial noise% = 0%, 10%, 20%, 30%, 40%, 50% added to Chowdary2006 and Chen2002.

examined. Also, it performs well on most of the other data sets. IFCE is relatively stable with varying parameter values and is robust to highly noisy synthetic data sets. Moreover, IFCE is computationally efficient.

Chapter 5

Conclusions and Future Work

In this dissertation, we have presented a new fuzzy cluster ensemble method IFCE. We have also evaluated IFCE through comparisons with numerous existing benchmark ensemble clustering and simple clustering methods on eight real cancer gene expression data sets. IFCE is the top performer on three of the eight data sets, more than any other methods examined. Also, it performs well on most of the other data sets. IFCE is relatively stable with varying parameter values and is robust to highly noisy synthetic data sets. Moreover, IFCE is computationally efficient.

For future work, we can extend our work in the following two perspectives: clustering algorithm and bioinformatics.

5.1 Conclusions

In the past few decades, the emerging breakthroughs in biotechnology such as microarray produced enormous amount of various types of large scaled biological data, including genomic data and gene expression data. As a result, bioinformatics faces a substantial challenge: how to extract meaningful knowledge from these data. One particular area of interest is extracting or finding cancer subtypes from gene expression data. Cancer subtyping can drastically improve patient treatment selection and outcome. Cancer subtyping can be done by clustering gene expression data samples. Various clustering algorithms have been applied, however most of them may not be effective to address two of the top challenges of noisy data and high-dimensional data

that are inherently characteristics of microarray experiments.

This dissertation proposes a noise robust fuzzy cluster ensemble IFCE to address the above two important challenges. IFCE uses the Modified Weighted Fuzzy Expected Value (MWFEV) method for computing cluster centers instead of the average or the mean method commonly used by other clustering approaches. The MWFEV method weighs outlier data objects less than more densely distributed data objects, and provides a more accurate center for a cluster [174]. Whereas, the average and the mean methods are sensitive to outliers, and may produce a less accurate and problematic cluster center. IFCE is a fuzzy or soft clustering algorithm. It uses fuzzy membership method in computing cluster centers and assigning feature vectors. In contrast, crisp or hard clustering uses Boolean method and are often highly sensitive to noise. While fuzzy clustering is more robust against noise thus pre-filtering of features can be avoided. This prevents the exclusion of domain objects relevant features for clustering and the data analysis [186]. IFCE is a cluster ensemble. It employs multiple base clusterings of IFCs to form a final consensus. Clustering ensemble improves robustness, stability and accuracy of clustering results. IFCE incorporates user interactivity functions. It provides users with the option of defining multiple parameters or thresholds. This way, users can optimize the clustering results of any particular data set. IFCE's time and space complexity converge to O(N). Whereas, most clustering and cluster ensemble algorithms have higher order of complexities especially regarding to the cost of time. IFCE's excellent efficiency in complexities makes it scalable and practical to very large scale data including high dimensional data.

We have conducted a comprehensive study on numerous data sets: the classical Iris data set, eight real cancer gene expression data sets, and ten synthetic noisy data sets. Real cancer gene expression data sets include cancers of breast, colon, prostate, blood, brain, bladder, and liver. During the experiments on eight real cancer gene expression data sets, we compared IFCE with ten existing benchmark ensemble clustering and simple clustering algorithms for CA validity measure, which

are: MULI-K, CCHC, GCC, CSPA, HGPA, MCLA, SL, CL, AL, and KM. The experimental results have shown that our algorithms are robust against noise and scalable to large data sets. More specifically, IFCE is the top performer on three of the eight data sets, more than any other methods examined. Also, it performs well on most of the other data sets. During the experiments on varying parameter and threshold values, the performance of IFCE is relatively stable. This proves IFCE to be a suitable and easy to use tool as it has relative consistent performance over various settings of important parameters or threshold. During the experiments on ten synthetic noisy data sets, IFCE is robust even to highly noisy data, the top challenges in cancer gene expression data.

Clustering cancer gene expression data is important. It is used in cancer subtyping and accurate cancer subtyping is crucial to treatment success. Traditional clinical data based cancer type diagnosis and treatment are based on the organ or tissue in which the cancer originates. A huge obstacle for effective treatments for cancers is that patients with similar cancer types respond differently to similar treatments. [108]. Cancer develops as a result of multiple genetic defects. Individuals with the same type of cancer often have dissimilar genetic defects in their tumors. Each patient should be treated according to the specific genetic defects in the tumor. Molecular data based cancer subtyping finds smaller groups or subtypes that a type of cancer can be divided into, based on certain characteristics of the cancer cells. These characteristics include DNA and/or other molecular changes of the cells. It is based on biomarkers including gene expression signatures. IFCE excels at addressing the challenges in clustering cancer gene expression data such as: noise, high dimensionality, accuracy and reliability, time complexity, and space complexity.

5.2 Future Work

We can extend our work in the following two perspectives for future work: clustering algorithm and bioinformatics. For clustering algorithm, we will discuss clustering and cluster ensemble. For bioinformatics, we will discuss incorporating biological

knowledge in clustering, additional biological data sets, biological based similarity measures, biological based validity measures, and time-series gene expression data clustering.

5.2.1 Clustering Algorithm

In the future, we could consider adopting different methods to compute cluster centers and different similarity measures or techniques to assign data objects to clusters. In the future, we could consider incorporating additional clustering algorithms to be the base clusterings of the ensemble. Additionally, we could experiment with different consensus functions in the ensemble including pairwise similarity-based consensus functions, graph-based consensus functions, mutual information-based consensus functions, and other types of voting-based consensus functions.

5.2.2 Bioinformatics

Clustering algorithms are unsupervised learning approaches, which means no prior knowledge is used. Fortunately, for gene expression data sets, some prior domain knowledge is often available. For example, some genes are known to be function-related. In the future, integrating such domain knowledge into the clustering process may improve the results substantially [132]. Gene Ontology (GO) can be used as prior domain knowledge to improve the performance of clustering methods [240]. GO is one of the rapidly increasing functional annotation resources. It is the framework for the model of biology. It defines concepts or classes used to describe gene function, and relationships between these concepts or classes. It classifies gene functions along three aspects: molecular function covers molecular activities of gene products, such as binding or catalysis; cellular component covers where gene products are active, such as the parts of a cell or its extracellular environment; biological process covers pathways and larger processes made up of the activities of multiple gene products, pertinent to the functioning of integrated living units such as cells, tissues, organs, and organisms [54]. GO annotations is the model of biology. Annotations are state-

ments describing the functions of specific genes, using concepts or classes in the Gene Ontology. Each statement is based on a specified piece of evidence. The simplest and most common annotation links one gene to one function [53].

There are many types of cancers and the amount of gene expression data sets generated for them are continuously increasing especially with new biotechnologies. The experimentation results of our IFCE look promising in the data sets we chose. In the future, we could experiment with additional data sets.

In addition, we could design or choose different similarity measures that are better suit for high-dimensional data as well as biological data. For example, Spearman correlation is more robust to noise than Pearson and Euclidean distance. LSS (Local Shape-based Similarity) [19] is based on the observation of biological relationships between genes [131]. The similarity measure [280] that uses an estimate of the GO-based similarity between two genes [32]. IBSA (Intrinsic Biological Separation Ability) method [131] employes semantic similarities among genes extracted from the GO, and uses the Best-Match Average of the Resnik measure [230, 219] as its biological proximity measure between genes.

Also, We could design or choose different validity measures that takes into account known biological knowledge, instead of simple external, internal, or relevant validity measures [124]. This way, the performance of IFCE is evaluated in terms of its ability to produce biologically meaningful clusters. For example, BSI (Biological Stability Index) [62] evaluates the stability of a clustering result through the removal of features [131]. BHI (Biological Homogeneity Index) [62] assesses the homogeneity through the biological terms extracted from the GO [131]. The validity measure that is based on the selection of relevant and non-redundant terms from GO [58, 131]. The validity measures that is based on semantic similarities from the GO [32, 131].

Biological processes are often dynamic, and researchers need to monitor them continuously or at multiple time points. Time-series gene expression data contain abundant information about such dynamic activity. Such data identify activated genes in a biological process, their rates of activity, their order, and their causal

effects. At any given time a cell only expresses a small fraction of all the genes in the organism's genome. Expressed genes reflect the cell's functional capacities and ability to respond to external stimuli. There are several categories of time-series experiments such as developmental processes, cyclic processes, and response to an external signal. Time-series gene expression data sets, including single-cell measurements and next-generation sequencing technologies, provide new opportunities while also raise new computational analysis challenges such as potential increased noise [21, 12]. In the future, we could apply IFCE to time-series gene expression data for capturing the temporal and multidimensional dynamics of complex cancer subtypes.

Bibliography

[1] Elke Achtert, Christian Böhm, Jörn David, Peer Kröger, and Arthur Zimek. Global correlation clustering based on the hough transform. *Statistical Analysis and Data Mining*, 1(3):111–127, 2008.

[2] Elke Achtert, Christian Bohm, Peer Kroger, and Arthur Zimek. Mining hierarchies of correlation clusters. In *Scientific and Statistical Database Management, 2006. 18th International Conference on*, pages 119–128. IEEE, 2006.

[3] Walaa A. Afifi and Hesham A. Hefny. Adaptive takagi-sugeno fuzzy model using weighted fuzzy expected value in wireless sensor network. In *Hybrid Intelligent Systems (HIS), 2014 14th International Conference on*, pages 225–231. IEEE, 2014.

[4] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 61–72, New York, NY, USA, 1999. ACM.

[5] Charu C. Aggarwal and Philip S. Yu. Finding generalized projected clusters in high dimensional spaces. *SIGMOD Rec.*, 29(2):70–81, May 2000.

[6] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2):94–105, June 1998.

[7] Shubham Agrawal, B. K. Panigrahi, and Manoj Kumar Tiwari. Multiobjective particle swarm algorithm with fuzzy clustering for electrical power dispatch. *IEEE Trans. Evolutionary Computation*, 12(5):529–541, 2008.

[8] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[9] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed

by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[10] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[11] Peter Andras. Kernel-kohonen networks. *Int. J. Neural Syst.*, 12(2):117–135, 2002.

[12] IP Androulakis, E Yang, and RR Almon. Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, 9:205–228, 2007.

[13] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, June 1999.

[14] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47, 2002.

[15] Roberto Avogadri and Giorgio Valentini. Fuzzy ensemble clustering based on random projections for dna microarray data analysis. *Artificial Intelligence in Medicine*, 45(2):173–183, 2009.

[16] J. Azimi, M. Mohammadi, A. Movaghar, and M. Analoui. Clustering ensembles using genetic algorithm. In *Computer Architecture for Machine Perception and Sensing, 2006. CAMP 2006. International Workshop on*, pages 119 –123, 18-20 2007.

[17] Shahariz Abdul Aziz and Jeyakody Parthiban. Fuzzy logic and its uses, 1996. URL `https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol2/jp6/article2.fuzzypeople.gif`. Accessed April 2018.

[18] Anthony J. Bagnall and Gareth J. Janacek. Clustering time series from ARMA models with clipped data. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *KDD*, pages 49–58. ACM, 2004.

[19] Rajarajeswari Balasubramaniyan. *Gene expression data analysis using novel methods: Predicting time delayed correlations and evolutionarily conserved functional modules*. PhD thesis, Philipps-Universität Marburg Fachbereich Biologie, 2005.

[20] GH Ball and DJ Hall. ISODATA, an iterative method of multivariate analysis and pattern classification. *IFIPS Congress*, 1965.

[21] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13:552 EP –, Jul 2012. Review Article.

[22] D. Barbara and P. Chen. Using the fractal dimension to cluster datasets. In *Proc. of the 6th International Conference on Knowledge Discovery and Data Mining*, pages 260–264. ACM Press, 2000.

[23] Daniel Barbara, Julia Couto, and Yi Li. COOLCAT: an entropy-based algorithm for categorical clustering. In *In Proceedings of the eleventh international conference on Information and knowledge management*, pages 582–589. ACM Press, 2002.

[24] Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: mining contrast sets. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 302–306, New York, NY, USA, 1999. ACM.

[25] Shai Ben-David and Nika Haghtalab. Clustering in the presence of background noise. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–280–II–288. JMLR.org, 2014.

[26] Pavel Berkhin. Survey Of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[27] J. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure I. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics*, 40(2):339–357, 1981.

[28] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13790–13795, November 2001.

[29] Christian Bhm, Karin Kailing, Hans-Peter Kriegel, and Peer Krger. Density connected clustering with local subspace preferences. In *ICDM*, pages 27–34. IEEE Computer Society, 2004.

[30] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, Aug. 2000.

[31] Cherie Blenkiron, Leonard D Goldstein, Natalie P Thorne, Inmaculada Spiteri, Suet-Feung Chin, Mark J Dunning, Nuno L Barbosa-Morais, Andrew E

Teschendorff, Andrew R Green, Ian O Ellis, et al. Microrna expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol*, 8(10):R214–R229, 2007.

[32] Nadia Bolshakova, Francisco Azuaje, and Pádraig Cunningham. Incorporating biological domain knowledge into cluster validity assessment. In *Workshops on Applications of Evolutionary Computation*, pages 13–22. Springer, 04 2006.

[33] Arianna Bottoni, Daniela Piccin, Federico Tagliati, Andrea Luchin, Maria Chiara Zatelli, and Ettore C Degli Uberti. miR-15a and miR-16-1 down-regulation in pituitary adenomas. *Journal of cellular physiology*, 204(1):280–285, 2005.

[34] Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. Batch kernel som and related laplacian methods for social network analysis. *Neurocomputing*, 71(7-9):1257–1273, 2008.

[35] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

[36] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, March 2004.

[37] D. Burton, J. Shore, and J. Buck. A generalization of isolated word recognition using vector quantization. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, volume 8, pages 1021–1024, Apr 1983.

[38] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich, and Carlo M. Croce. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*, 99(24):15524–15529, 2002.

[39] Francesco Camastra and Alessandro Verri. A novel kernel method for clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):801–804, 2005.

[40] Soumen Chakrabarti. *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann, 2003.

[41] Jennifer A. Chan, Anna M. Krichevsky, and Kenneth S. Kosik. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer research*, 65(14):6029–6033, 2005.

[42] Peter Cheeseman and John Stutz. Advances in knowledge discovery and data mining. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth,

and Ramasamy Uthurusamy, editors, *Bayesian classification (AutoClass): theory and results*, pages 153–180. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[43] Jiun-Rung Chen. *Efficient Biclustering Methods for Microarray Databases*. PhD thesis, National Sun Yat-sen University, 2010.

[44] X. Chen, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K. M. Lai, J. Ji, S. Dudoit, I. O. Ng, M. Van De Rijn, D. Botstein, and P. O. Brown. Gene expression patterns in human liver cancers. *Molecular biology of the cell*, 13(6):1929–1939, 2002.

[45] Yonghui Chen, Kevin D Reilly, Alan P Sprague, and Zhijie Guan. SEQOPTICS: a protein sequence clustering system. *BMC bioinformatics*, 7(Suppl 4):S10, 2006.

[46] Chun Hung Cheng, Ada Wai-Chee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In Usama M. Fayyad, Surajit Chaudhuri, and David Madigan, editors, *KDD*, pages 84–93. ACM, 1999.

[47] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proc. of the 8th Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, 2000.

[48] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The journal of molecular diagnostics*, 8(1):31–39, February 2006.

[49] Itay Chowers, Dongmei Liu, Ronald H Farkas, Tushara L Gunatilaka, Abigail S Hackam, Steven L Bernstein, Peter A Campochiaro, Giovanni Parmigiani, and Donald J Zack. Gene expression variation in the adult human retina. *Human molecular genetics*, 12(22):2881–2893, 2003.

[50] Itay Chowers, Dongmei Liu, Ronald H Farkas, Tushara L Gunatilaka, Abigail S Hackam, Steven L Bernstein, Peter A Campochiaro, Giovanni Parmigiani, and Donald J Zack. Gene expression variation in the adult human retina. *Human molecular genetics*, 12(22):2881–2893, 2003.

[51] SA Ciafre, S Galardi, A Mangiola, M Ferracin, C-G Liu, G Sabatino, M Negrini, G Maira, CM Croce, and MG Farace. Extensive modulation of a set of micrornas in primary glioblastoma. *Biochemical and biophysical research communications*, 334(4):1351–1358, 2005.

[52] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[53] Gene Ontology Consortium. Annotations, 1999. URL `http://www.geneontology.org/`. Accessed April 2018.

[54] Gene Ontology Consortium. Ontology, 1999. URL `http://www.geneontology.org/`. Accessed April 2018.

[55] Robson Leonardo Ferreira Cordeiro, Agma JM Traina, and Christos Faloutsos. Finding clusters in subspaces of very large, multi-dimensional datasets. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 625–636. IEEE, 2010.

[56] David W Corne, Nick R Jerram, Joshua D Knowles, Martin J Oates, et al. Pesa-ii: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO2001*, 2001.

[57] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine learning*, 20(3):273–297, September 1995.

[58] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007.

[59] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, March 2000.

[60] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matran. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25, 1997.

[61] R. DAndrade. U-statistic hierarchical clustering. *Psychometrika*, 4, 1978.

[62] Susmita Datta and Somnath Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC bioinformatics*, 7(1):397, 2006.

[63] Michele De Laurentiis, Grazia Arpino, Erminia Massarelli, Angela Ruggiero, Chiara Carlomagno, Fortunato Ciardiello, Giampaolo Tortora, Diego D'Agostino, Francesca Caputo, Giuseppe Cancello, et al. A meta-analysis on the interaction between her-2 expression and response to endocrine treatment in advanced breast cancer. *Clinical Cancer Research*, 11(13):4741–4748, 2005.

[64] Marcilio De Souto, Ivan Costa, Daniel de Araujo, Teresa Ludermir, and Alexander Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):1–14, 2008. 10.1186/1471-2105-9-497.

[65] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, 2002.

[66] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal (British Computer Society)*, 20(4):364366, 1977.

[67] Doulaye Dembele and Philippe Kastner. Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8):973–980, 2003.

[68] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[69] P D'haeseleer. How does gene expression clustering work? *Nature biotechnology*, 23(12):1499–1501, 2005.

[70] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM, 2003.

[71] C. Ding and Xiaofeng He. Cluster merging and splitting in hierarchical clustering algorithms. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 139–146, 2002.

[72] Marco Dorigo and Thomas Stützle. The ant colony optimization metaheuristic: Algorithms, applications, and advances. In *Handbook of metaheuristics*, pages 250–285. Springer, 2003.

[73] K.-L. Du. Clustering: A neural network approach. *Neural Networks*, 23(1):89–107, 2010.

[74] Zhihua Du, Yiwei Wang, and Zhen Ji. PK-means: A new algorithm for gene clustering. *Computational Biology and Chemistry*, 32(4):243–247, 2008.

[75] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001.

[76] Sandrine Dudoit and Jane Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[77] Joseph C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[78] L. Dyrskjot, T. Thykjaer, M. Kruh0ffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Orntoft. Identifying distinct classes of bladder carcinoma using microarrays. *Nature genetics*, 33(1):90–96, 2002.

[79] M. B. Eisen and P. O. Brown. DNA arrays for analysis of gene expression. *Methods Enzymol*, 303:179–205, 1999.

[80] A. Enright and C. Ouzounis. GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 16(5):451–457, 2000.

[81] S. Eschrich, Jingwei Ke, L. O. Hall, and D. B. Goldgof. Fast accurate fuzzy clustering through data reduction. *Fuzzy Systems, IEEE Transactions on*, 11(2), Apr 2003.

[82] Martin Ester, Hans P. Kriegel, Jorg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[83] F.T. Evers, F. Höppner, F. Klawonn, J.C. Rush, R. Kruse, I. Berdrow, and T. Runkler. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Jossey-Bass higher and adult education series. Wiley, 1999.

[84] A. Ferligoj and V. Batagelj. Direct multicriterion clustering algorithms. *Journal of Classification*, 9:43–61, 1992.

[85] Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In Tom Fawcett and Nina Mishra, editors, *International Conference on Machine Learning*, pages 186–193. AAAI Press, 2003.

[86] Xiaoli Zhang Fern and Carla E Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, page 36. ACM, 2004.

[87] David B Fogel. *Evolutionary computation: toward a new philosophy of machine intelligence*. John Wiley & Sons, 2006.

[88] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[89] Ana L. N. Fred. Finding consistent clusters in data partitions. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 309–318. Springer, 2001.

[90] Ana L. N. Fred and Anil K. Jain. Data clustering using evidence accumulation. In *Proceedings 16th International Conference on Pattern Recognition*, volume 4, pages 276–280. IEEE, 2002.

[91] Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72. ACM Press, 1999.

[92] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering - theory, algorithms, and applications.* SIAM, 2007.

[93] Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. CACTUS clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 73–83, New York, NY, USA, 1999. ACM.

[94] Mitchell E. Garber, Olga G. Troyanskaya, Karsten Schluens, Simone Petersen, Zsuzsanna Thaesler, Manuela Pacyna-Gengelbach, Matt van de Rijn, Glenn D. Rosen, Charles M. Perou, Richard I. Whyte, Russ B. Altman, Patrick O. Brown, David Botstein, and Iver Petersen. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences*, 98(24):1378413789, November 2001.

[95] Luis A. Garca-Escudero, Alfonso Gordaliza, Carlos Matrn, and Agustin Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36(3):1324–1345, 06 2008.

[96] A. Gersho and B. Ramamurthi. Image coding using vector quantization. *International Conference on Acoustics, Speech, and Signal Processing*, 1:428–431, April 1982.

[97] Vito Di Ges, Raffaele Giancarlo, Giosu Lo Bosco, Alessandra Raimondi, and Davide Scaturro. GenClust: A genetic algorithm for clustering gene expression data. *BMC Bioinformatics*, 6(1):289–289, 2005.

[98] Gad Getz, Erel Levine, and Eytan Domany. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, 97(22):12079–12084, 2000.

[99] Reza Ghaemi, Md. Nasir Sulaiman, Hamidah Ibrahim, and Norwati Mustapha. A survey: Clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, 50:636–645, 2009.

[100] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Clustering categorical data: an approach based on dynamical systems. *The International Journal on Very Large Data Bases*, 8(3-4):222–236, February 2000.

[101] M. Girolami. Mercer kernel-based clustering in feature space. *Neural Networks, IEEE Transactions on*, 13(3):780–784, August 2002.

[102] Github. Golub (1999). URL `https://github.com/ramhiser/datamicroarray/wiki/Golub-(1999)`. Accessed April 2018.

[103] Sanjay Goil, Harsha Nagesh, and Alok Choudhary. MAFIA: Efficient and scalable subspace clustering for very large data sets. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 443–452, 1999.

[104] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[105] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64, 1969.

[106] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: an efficient clustering algorithm for large databases. *SIGMOD Record*, 27(2):73–84, June 1998.

[107] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345 – 366, 2000.

[108] Sudheer Gupta, Kumardeep Chaudhary, Rahul Kumar, Ankur Gautam, Jagpreet Singh Nanda, Sandeep Kumar Dhanda, Samir Kumar Brahmachari, and Gajendra P. S. Raghava. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Scientific Reports*, 6(23857), 2016.

[109] Valerie Guralnik and George Karypis. A scalable algorithm for clustering sequential data. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, *ICDM*, pages 179–186. IEEE Computer Society, 2001.

[110] Dan Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press, 1997.

[111] Julia Handl and Joshua Knowles. Evolutionary multiobjective clustering. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 1081–1091. Springer, 2004.

[112] Julia Handl and Joshua Knowles. An evolutionary approach to multiobjective clustering. *Evolutionary Computation, IEEE Transactions on*, 11(1):56–76, 2007.

[113] Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[114] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, Oct 1990.

[115] Pierre Hansen and Brigitte Jaumard. Cluster analysis and mathematical programming. *Mathematical programming*, 79(1-3):191–215, 1997.

[116] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100108, 1979.

[117] J.A. Hartigan. Clustering algorithms. In *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[118] Aboul-Ella Hassanien, Mariofanna G Milanova, Tomasz G Smolinski, and Ajith Abraham. Computational intelligence in solving bioinformatics problems: Reviews, perspectives, and challenges. In *Computational Intelligence in Biomedicine and Bioinformatics*, pages 3–47. Springer, 2008.

[119] Michael J Heller. Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

[120] Alexander Hinneburg and Daniel A. Keim. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *Proceedings of the 25th International Conference on Very Large Data Bases*, pages 506–517. Morgan Kaufmann, 1999.

[121] J. Holland. *Adaption in Natural and Artificial Systems*. Ann Arbor, MI: Univ. Michigan Press, 1975.

[122] Jun Hou, Joachim Aerts, Bianca Den Hamer, Wilfred Van Ijcken, Michael Den Bakker, Peter Riegman, Cor van der Leest, Peter van der Spek, John A Foekens, Henk C Hoogsteden, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one*, 5(4):e10312, 2010.

[123] P. V. C. Hough. Methods and means for recognizing complex patterns, December 1962. US Patent 3069654.

[124] Natthakan Iam-on, Tossapon Boongoen, and Simon Garrett. LCE: a link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, 26(12):1513–1519, 2010.

[125] V. Ilango, R. Subramanian, and V. Vasudevan. Cluster analysis research design model, problems, issues, challenges, trends and tools. *International Journal on Computer Science and Engineering*, 3(8):2926–2934, 2011.

[126] Marilena V Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, et al. Microrna gene expression deregulation in human breast cancer. *Cancer research*, 65(16):7065–7070, 2005.

[127] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[128] Anil K. Jain. Data clustering: User's dilemma. In *Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '07, pages 1–1, Berlin, Heidelberg, 2007. Springer-Verlag.

[129] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[130] J.S.R. Jang, C.T. Sun, and E. Mizutani. *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. MATLAB curriculum series. Prentice Hall, 1997.

[131] Pablo Andretta Jaskowiak. *On the evaluation of clustering results: measures, ensembles, and gene expression data analysis*. PhD thesis, Universidade de São Paulo, 2015.

[132] Daxin Jiang. Mining coherent patterns and clusters from genomic data, May 2005.

[133] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 16(11):1370–1386, November 2004.

[134] Steven M Johnson, Helge Grosshans, Jaclyn Shingara, Mike Byrom, Rich Jarvis, Angie Cheng, Emmanuel Labourier, Kristy L Reinert, David Brown, and Frank J Slack. RAS Is Regulated by the let-7 MicroRNA Family. *Cell*, 120(5):635–647, 2005.

[135] Susan Jones, David TA Daley, Nicholas M Luscombe, Helen M Berman, and Janet M Thornton. Protein–rna interactions: a structural analysis. *Nucleic acids research*, 29(4):943–954, 2001.

[136] Cengiz Kahraman, Basar ztaysi, and Sezi evik Onar. A comprehensive literature review of 50 years of fuzzy set theory. *International Journal of Computational Intelligence Systems*, 9(sup1):3–24, 2016.

[137] Mostafa Karbasi, Zeeshan Bhatti, Reza Aghababaeyan, Sara Bilal, Abdolvahab Ehsani Rad, Asadullah Shah, and Ahmad Waqas. Real-time hand detection by depth images: A survey. *Jurnal Teknologi*, 78, 02 2016.

[138] Yoko Karube, Hisaaki Tanaka, Hirotaka Osada, Shuta Tomida, Yoshio Tatematsu, Kiyoshi Yanagisawa, Yasushi Yatabe, Junichi Takamizawa, Shinichiro Miyoshi, Tetsuya Mitsudomi, et al. Reduced expression of dicer associated with poor prognosis in lung cancer patients. *Cancer science*, 96(2):111–115, 2005.

[139] G. Karypis, R. Aggarwal, V. Kumar, and Shashi Shekhar. Multilevel hypergraph partitioning: applications in vlsi domain. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 7(1):69–79, March 1999.

[140] G. Karypis, E. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, Aug. 1999.

[141] George Karypis and Vipin Kumar. *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices.* University of Minnesota, Department of Computer Science, September 1998.

[142] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[143] U. Kaymak and M. Setnes. Fuzzy clustering with volume prototypes and adaptive cluster merging. *IEEE Transactions on Fuzzy Systems*, 10(6):705–712, Dec 2002.

[144] Paul Kellam, Xiaohui Liu, Nigel Martin, Christine Orengo, Stephen Swift, and Allan Tucker. Comparing, contrasting and combining clusters in viral gene expression data. In *Proceedings of 6th workshop on intelligent data analysis in medicine and pharmocology*, pages 56–62, 2001.

[145] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(4):580–585, July 1985.

[146] James F Kennedy, James Kennedy, and Russell C Eberhart. *Swarm intelligence*. Morgan Kaufmann, 2001.

[147] Javed Khan, Jun S. Wei, Markus Ringnr, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673679, June 2001.

[148] Eun-Youn Kim, Seon-Young Kim, Daniel Ashlock, and Dougu Nam. Multi-k: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC bioinformatics*, 10(1):260, 2009.

[149] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[150] Teuvo Kohonen. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[151] Dorota T Kopycka-Kedzierawski and Ronald J Billings. A longitudinal study of caries onset in initially caries-free children and baseline salivary mutans streptococci levels: a kaplan–meier survival analysis. *Community dentistry and oral epidemiology*, 32(3):201–209, 2004.

[152] Peer Krger, Hans-Peter Kriegel, and Karin Kailing. Density-connected subspace clustering for high-dimensional data. In Michael W. Berry, Umeshwar Dayal, Chandrika Kamath, and David B. Skillicorn, editors, *SDM*, volume 4. SIAM, 2004.

[153] Hans-Peter Kriegel, Peer Krger, Matthias Renz, and Sebastian Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In *Data Mining, Fifth IEEE International Conference on*, pages 250–257. IEEE Computer Society, 2005.

[154] R. Krishnapuram, O. Nasraoui, and H. Frigui. The fuzzy c spherical shells algorithm: A new approach. *IEEE Transactions on Neural Networks*, 3(5):663–671, Sep 1992.

[155] N Kumar and RS Joshi. Data clustering using artificial neural networks. In *Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007)*, pages 197–200, 2007.

[156] L.I. Kuncheva and S.T. Hadjitodorov. Using diversity in cluster ensembles. *IEEE International Conference on Systems, Man & Cybernetics*, 2:1214–1219, 2004.

[157] L.I. Kuncheva and D.P. Vetrov. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1798 –1808, nov. 2006.

[158] P. Laiho, A. Kokko, S. Vanharanta, R. Salovaara, H. Sammalkorpi, H. Jarvinen, J. P. Mecklin, T. J. Karttunen, K. Tuppurainen, V. Davalos, S. Schwartz, D. Arango, M. J. Makinen, and L. A. Aaltonen. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*, 26(2):312–320, July 2006.

[159] Jacques Lapointe, Chunde Li, John P. Higgins, Matt van de Rijn, Eric Bair, Kelli Montgomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, Peter Ekman, Angelo M. DeMarzo, Robert Tibshirani, David Botstein, Patrick O. Brown, James D. Brooks, and Jonathan R. Pollack. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):811–816, January 2004.

[160] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, pages 86–112, 2006.

[161] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2000.

[162] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843 – 854, 1993.

[163] Zhengdeng Lei, Iain Beehuat Tan, Kakoli Das, Niantao Deng, Hermioni Zouridis, Sharon Pattison, Clarinda Chua, Zhu Feng, Yeoh Khay Guan, Chia Huey Ooi, et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology*, 145(3):554–565, 2013.

[164] H. Li, K. Zhang, and T. Jiang. Minimum entropy clustering and applications to gene expression analysis. *Proceedings IEEE Computational Systems Bioinformatics Conference*, pages 142–151, 2004.

[165] Li Li, Chang Liu, Fang Wang, Wei Miao, Jie Zhang, Zhiqian Kang, Yihan Chen, and Luying Peng. Unraveling the hidden heterogeneities of breast cancer based on functional miRNA cluster. *PLOS ONE*, 9(1):e87601–v, 2014.

[166] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, Oct. 2004.

[167] Y. Liang, M. Diehn, N. Watson, A. W. Bollen, K. D. Aldape, M. K. Nicholas, K. R. Lamborn, M. S. Berger, D. Botstein, P. O. Brown, and M. A. Israel. Gene expression profiling reveals molecularly and clinically distinct subtypes

of glioblastoma multiforme. *Proceedings of the National Academy of Sciences USA*, 102(16):5814–5819, 2005.

[168] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature genetics*, 21(1 Suppl):20–4, Jan. 1999.

[169] Jingwei Liu and Meizhi Xu. Kernelized fuzzy attribute C-means clustering algorithm. *Fuzzy Sets and Systems*, 159(18):2428–2445, 2008.

[170] L. Liu, D. M. Hawkins, S. Ghosh, and S. S. Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences USA*, 100(23):13167–13172, 2003.

[171] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 1982.

[172] Eric F. Lock and David B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.

[173] Carl G. Looney. *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*. Oxford University Press, 1997.

[174] Carl G. Looney. Interactive clustering and merging with a new fuzzy expected value. *Pattern Recognition*, 35(11):2413 – 2423, 2002.

[175] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, et al. Microrna expression profiles classify human cancers. *nature*, 435(7043):834–838, 2005.

[176] Huilan Luo, Furong Jing, and Xiaobing Xie. Combining multiple clusterings using information theory based genetic algorithm. In *Computational Intelligence and Security, 2006 International Conference on*, volume 1, pages 84–89, Nov 2006.

[177] P.C.H. Ma, K.C.C. Chan, Xin Yao, and D.K.Y. Chiu. An evolutionary clustering algorithm for gene expression microarray data analysis. *Evolutionary Computation, IEEE Transactions on*, 10(3):296 – 314, june 2006.

[178] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.*, 1(14):281297, 1967.

[179] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45, 2004.

[180] S. W. Mahfoud. *Niching Methods for Genetic Algorithms*. PhD thesis, University of Illinois at Urbana-Champaign, 1995.

[181] Satyendra Nath Mandal, J Pal Choudhury, and SR Bhadra Chaudhuri. In search of suitable fuzzy membership function in prediction of time series data. *International Journal of Computer Science Issues*, 9(3):293–302, 2012.

[182] Andrew J Maniotis, Robert Folberg, Angela Hess, Elisabeth A Seftor, Lynn MG Gardner, Jacob Pe'er, Jeffrey M Trent, Paul S Meltzer, and Mary JC Hendrix. Vascular channel formation by human melanoma cells *in Vivo* and *in Vitro*: Vasculogenic mimicry. *The American journal of pathology*, 155(3):739–752, 1999.

[183] Elena Marchiori and Jason H Moore. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 5th European Conference, EvoBIO 2007, Valencia, Spain, April 11-13, 2007, Proceedings*, volume 4447. Springer Science & Business Media, 2007.

[184] Shawn Martin. Machine learning based bioinformatics algorithms: Application to chemicals, 2010. URL `http://www.cs.otago.ac.nz/homepages/smartin/publications_long.php`. Accessed April 2018.

[185] Thomas M. Martinetz and Klaus J. Schulten. A "neural gas" network learns topologies. In Teuvo Kohonen, Kai Mäkisara, Olli Simula, and Jari Kangas, editors, *Proceedings of the International Conference on Artificial Neural Networks* (Espoo, Finland), pages 397–402. Amsterdam; New York: North-Holland, 1991.

[186] Bronwyn Carlisle Matthias Futschik. Noise-robust soft clustering of gene expression time-course data. *Journal of Bioinformatics and Computational Biology*, 3(4), 2005.

[187] Michael D Mattie, Christopher C Benz, Jessica Bowers, Kelly Sensinger, Linda Wong, Gary K Scott, Vita Fedele, David Ginzinger, Robert Getts, and Chris Haqq. Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. *Molecular cancer*, 5(1):1–24, 2006.

[188] G. J. McLachlan and K. E. Basford. *Mixture models: inference and applications to clustering*. Marcel Dekker Inc, New York / Basel, 1988.

[189] G. J. McLachlan, R. W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

[190] Geoffrey J. McLachlan, Kim-Anh Do, and Christophe Ambroise. *Microarrays in Gene Expression Studies*, pages 1–29. John Wiley & Sons, Inc., 2005.

[191] Muhammad Aamer Mehmood, Ujala Sehar, and Niaz Ahmad. Bioinformatics applications, 2014. URL `https://www.omicsonline.org/articles-images/data-mining-genomics-Application-bioinformatics-tools-5-158-g001.png`. Accessed April 2018.

[192] Eric Melse. Spectramap biplot iris flower data set, 2010. URL `https://commons.wikimedia.org/wiki/File:Spectramap_Biplot_Iris_Flower_Data_Set_FULL.jpg`. Accessed April 2018.

[193] Markus Metzler, Monika Wilda, Kerstin Busch, Susanne Viehmann, and Arndt Borkhardt. High expression of precursor microRNA-155/BIC RNA in children

with Burkitt lymphoma. *Genes, Chromosomes and Cancer*, 39(2):167–169, 2004.

[194] Michael Z Michael, Susan M O'Connor, Nicholas G van Holst Pellekaan, Graeme P Young, and Robert J James. Reduced accumulation of specific microRNAs in Colorectal Neoplasia. *Molecular Cancer Research*, 1(12):882–891, 2003.

[195] Mithun Mitra. *Dissecting the nucleic acid chaperone properties of retroviral nucleocapsid proteins*. University of Minnesota, 2007.

[196] Gabriela Moise, Jrg Sander, and Martin Ester. P3C: A robust projected clustering algorithm. In *ICDM*, pages 414–425. IEEE Computer Society, 2006.

[197] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.

[198] Tadeusz Morzy, Marek Wojciechowski, and Maciej Zakrzewicz. Pattern-oriented hierarchical clustering. In *Proceedings of the third East-European Symposium on Advances in Databases and Information Systems ADBIS99, Slovenia, LNCS 1691*, pages 179–190, 1999.

[199] Anirban Mukhopadhyay and Ujjwal Maulik. A multiobjective approach to MR brain image segmentation. *Applied Soft Computing*, 11(1):872–880, 2011.

[200] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[201] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, March 1970.

[202] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

[203] Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 144–155, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[204] Luis ngel Garca-Escudero and Alfonso Gordaliza. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969, 1999.

[205] Hung Nguyen, Sangam Shrestha, Sorin Draghici, and Tin Nguyen. Pinsplus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, page bty1049, 2018.

[206] N. Nguyen and R. Caruana. Consensus clusterings. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 607–612, Oct 2007.

[207] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici. A novel approach for data integration and disease subtyping. *Genome research*, 27(12):2025–2039, 2017.

[208] Tin Nguyen, Rebecca Tagett, Diana Diaz, and Sorin Draghici. A novel approach for data integration and disease subtyping. *Genome research*, pages gr–215129, 2017.

[209] Sarfaraz K Niazi. *Handbook of bioequivalence testing*. CRC Press, 2014.

[210] Catherine L Nutt, DR Mani, Rebecca A Betensky, Pablo Tamayo, J Gregory Cairncross, Christine Ladd, Ute Pohl, Christian Hartmann, Margaret E McLaughlin, Tracy T Batchelor, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research*, 63(7):1602–1607, 2003.

[211] Ann L. Oberg, Amy J. French, Aaron L. Sarver, Subbaya Subramanian, Bruce W. Morlan, Shaun M. Riska, Pedro M. Borralho, Julie M. Cunningham, Lisa A. Boardman, Liang Wang, Thomas C. Smyrk, Yan Asmann, Clifford J. Steer, and Stephen N. Thibodeau. miRNA expression in colon polyps provides evidence for a multihit model of colon cancer. *PLoS ONE*, 6(6):e20465–e20465, 06 2011.

[212] Elizabeth O'Day and Ashish Lal. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Research*, 12(2):201–201, 2010.

[213] Oluwasogo Oluwafemi Ogundowole. Basic analysis of the iris data set using python, 2017. URL `https://medium.com/codebagng/basic-analysis-of-the-iris-data-set-using-python-2995618a6342`. Accessed April 2018.

[214] Zdzisław Pawlak. Elementary rough set granules: Toward a rough set. *Rough-Neural Computing: Techniques for Computing with Words*, page 5, 2012.

[215] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[216] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA*, 85(8):2444–2448, April 1988.

[217] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

[218] Charles M Perou, Stefanie S Jeffrey, Matt Van De Rijn, Christian A Rees, Michael B Eisen, Douglas T Ross, Alexander Pergamenschikov, Cheryl F Williams, Shirley X Zhu, Jeffrey CF Lee, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217, 1999.

[219] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(5):S4, 2008.

[220] Harun Pirim, Dilip Gautam, Tanmay Bhowmik, Andy D. Perkins, and Burak Ekioglu. Performance of an ensemble clustering algorithm on biological data sets. *Mathematical and Computational Applications*, 16(1):87–96, 2011.

[221] Clara Pizzuti and Domenico Talia. P-AutoClass: Scalable parallel clustering for mining large data sets. *Knowledge and Data Engineering, IEEE Transactions on*, 15(3):629–641, March 2003.

[222] Scott L Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M Sturla, Michael Angelo, Margaret E McLaughlin, John YH Kim, Liliana C Goumnerova, Peter M Black, Ching Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[223] Kati P Porkka, Minja J Pfeiffer, Kati K Waltering, Robert L Vessella, Teuvo LJ Tammela, and Tapio Visakorpi. MicroRNA expression profiling in prostate cancer. *Cancer research*, 67(13):6130–6135, 2007.

[224] ArcGIS Pro. Applying fuzzy logic to overlay rasters. URL `http://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/GUID-DCA2552E-5EE0-449C-A7E3-6F9CBFE97BC8-web.png`. Accessed April 2018.

[225] Cecilia Magdalena Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. A Monte Carlo algorithm for fast projective clustering. In Michael J. Franklin, Bongki Moon, and Anastassia Ailamaki, editors, *SIGMOD Conference*, pages 418–427. ACM, 2002.

[226] J. Quackenbush. Computational analysis of cDNA microarray data. *Nature Reviews*, 6(2):418–428, 2001.

[227] Sridhar Ramaswamy and Todd R. Golub. Dna microarrays in clinical oncology. *Journal of Clinical Oncology*, 20(7):1932–1941, 2002.

[228] Marco Ramoni, Paola Sebastiani, and Paul R. Cohen. Bayesian clustering by dynamics. *Machine Learning*, 47(1):91–121, 2002.

[229] ResearchHubs. Fuzzy logic membership function, 2015. URL `http://researchhubs.com/post/engineering/fuzzy-system/fuzzy-membership-function.html`. Accessed April 2018.

[230] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130, 1999.

[231] Roco Romero-Zliz, Cristina Rubio-Escudero, J. P. Cobb, Francisco Herrera, Oscar Cordn, and Igor Zwir. A multiobjective evolutionary conceptual clustering methodology for gene annotation within structural databases: A case of study on the gene ontology database. *Evolutionary Computation, IEEE Transactions on*, 12(6):679–701, 2008.

[232] Debasis Samanta. Fuzzy membership functions. URL `http://www.nid.iitkgp.ernet.in/dsamanta/courses/archive/sca/Archives/Chapter\%203\%20Fuzzy\%20Membership\%20Functions.pdf`. Accessed April 2018.

[233] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data mining and knowledge discovery*, 2(2):169–194, June 1998.

[234] E. Schikuta. Grid-clustering: an efficient hierarchical clustering method for very large data sets. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, 2:101–105, August 1996.

[235] Erich Schikuta and Martin Erhart. The BANG-clustering system: Grid-based data analysis. In *Advances in Intelligent Data Analysis Reasoning about Data*, pages 513–524. Springer, 1997.

[236] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik. Incorporating invariances in support vector learning machines. In *Artificial Neural NetworksICANN 96*, pages 47–52. Springer, 1996.

[237] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, July 1998.

[238] Ruty Shai, Tao Shi, Thomas J. Kremen, Steve Horvath, Linda M. Liau, Timothy F. Cloughesy, Paul S. Mischel, and Stanley F. Nelson. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, 22(4918), 2003.

[239] Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 428–439. Morgan Kaufmann Publishers Inc., 1998.

[240] Yijing Shen, Wei Sun, and Ker-Chau Li. Dynamically weighted clustering with noise set. *Bioinformatics*, Vol. 26 no. 3, 2010.

[241] Weiguo Sheng, Allan Tucker, and Xiaohui Liu. *Clustering with Niching Genetic K-means Algorithm*, pages 162–173. Lecture Notes in Computer Science, Genetic and Evolutionary Computation – GECCO 2004. Springer Berlin / Heidelberg, 2004.

[242] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, 10(12):1–20, 12 2015.

[243] R. Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society)*, 16(1):30–34, 1973.

[244] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, March 2002.

[245] SlidePlayer. Fuzzy logic. URL `http://slideplayer.com/slide/4959737/16/images/3/Fuzzy+Logic+Example.jpg`. Accessed April 2018.

[246] Donna K Slonim, Pablo Tamayo, Jill P Mesirov, Todd R Golub, and Eric S Lander. Class prediction and discovery using gene expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 263–272. ACM, 2000.

[247] Justine R. Smith. Bioinformatics and the eye. *Journal of Ocular Biology, Diseases, and Informatics*, 2(4):161–163, Dec 2009.

[248] Temple F. Smith and Michael S. Waterman. New stratigraphic correlation techniques. *The Journal of Geology*, 88(4):451–457, Jul. 1980.

[249] Tomasz G Smolinski, Mariofanna G Milanova, and Aboul-Ella Hassanien. *Computational Intelligence in Biomedicine and Bioinformatics: Current trends and applications*, volume 151. Springer, 2009.

[250] Padhraic Smyth. Clustering sequences with hidden markov models. In Michael Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing*, pages 648–654. MIT Press, 1996.

[251] P. Sneath and R. Sokal. Numerical taxonomy. In *Numerical Taxonomy*. W.H. Freeman and Company, 1973.

[252] The American Cancer Society and LIVESTRONG Launch First Global Economic Cost of Cancer Report. Cancer, 2013. URL `http://pressroom.cancer.org/index.php?s=43\&item=262`. Accessed Oct. 2013.

[253] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin*, 38:1409–1438, 1958.

[254] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M. Perou, Per E. Lønning, Patrick O. Brown, Anne-Lise Børresen-Dale, and David Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003.

[255] SpectraWorks. Data and noise, 2013. URL `http://www.spectraworks.com/Help/lowsignal.html`. Accessed April 2018.

[256] SpectraWorks. Data and signal, 2013. URL `http://www.spectraworks.com/Help/lowsignal.html`. Accessed April 2018.

[257] Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, March 2005.

[258] H. Steinhaus. Sur la division des corp materiels en parties. *Bulletin of Acad. Polon. Sci.*, 4(12):801–804, 1956.

[259] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[260] Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow, and Donald Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.

[261] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101(9):2981–2986, 2004.

[262] Chun Tang and Aidong Zhang. An iterative strategy for pattern discovery in high-dimensional data sets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 10–17. ACM, 2002.

[263] Chun Tang, Li Zhang, Aidong Zhang, and Murali Ramanathan. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*, pages 41–48, 2001.

[264] G. Teng, C. He, J. Xiao, Y. He, B. Zhu, and X. Jiang. Cluster ensemble framework based on the group method of data handling. *Appl. Soft Comput.*, 43(C):35–46, June 2016.

[265] Alexander P. Topchy, Anil K. Jain, and William F. Punch. Combining multiple weak clusterings. In *Proceedings of the IEEE International Conference on Data Mining*, pages 331–338. IEEE Computer Society, 2003.

[266] Alexander P Topchy, Anil K Jain, and William F Punch. A mixture model for clustering ensembles. In *Proceedings SIAM International Conference on Data Mining*. SIAM, 2004.

[267] Boston University. Bioinformatics and related disciplines, 2014. URL `http://www.bu.edu/bioinformatics/files/2009/04/bioinformatics-departments.png`. Accessed April 2018.

[268] Upatras. A simple example of principal component analysis on the build in iris dataset. URL `https://eclass.upatras.gr/modules/document/file.php/ECON1332/PCA.R.txt`. Accessed April 2018.

[269] Laura J. Van't Veer and René Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–570, April 2008.

[270] Sandro Vega-Pons and José Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(3):337–372, 2011.

[271] Michail Vlachos, Jessica Lin, Eamonn Keogh, and Dimitrios Gunopulos. A Wavelet-Based anytime algorithm for k-means clustering of time series. In *In Proceedings Workshop on Clustering High Dimensionality Data and Its Applications*, pages 23–30, 2003.

[272] Haiying Wang, Huiru Zheng, and Francisco Azuaje. Poisson-based self-organizing feature maps and hierarchical clustering for serial analysis of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2):163–175, 2007.

[273] Wei Wang, Jiong Yang, and Richard R. Muntz. STING: A statistical information grid approach to spatial data mining. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 186–195. Morgan Kaufmann, 1997.

[274] Wei Wang, Jiong Yang, and Richard R. Muntz. STING+: An approach to active spatial data mining. In Masaru Kitsuregawa, Michael P. Papazoglou, and Calton Pu, editors, *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 116–125. IEEE Computer Society, 1999.

[275] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[276] WHO. Cancer, 2013. URL `http://www.who.int/mediacentre/factsheets/fs297/en/`. Accessed Oct. 2013.

[277] Matthew D Wilkerson and D Neil Hayes. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.

[278] Matthew D Wilkerson, Xiaoying Yin, Katherine A Hoadley, Yufeng Liu, Michele C Hayward, Christopher R Cabanski, Kenneth Muldrew, C Ryan Miller, Scott H Randell, Mark A Socinski, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research*, 16(19):4864–4875, 2010.

[279] Junjie Wu. *K-means Based Consensus Clustering*, pages 155–175. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[280] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[281] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, Aug 1991.

[282] Eric P. Xing and Richard M. Karp. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, 17(suppl 1):S306–S315, 2001.

[283] Yimin Xiong and Dit-Yan Yeung. Time series clustering with arma mixtures. *Pattern Recognition*, 37(8):1675–1689, 2004.

[284] R. Xu, G. C. Anagnostopoulos, and D. C. Wunsch. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(1):65–77, Jan-Mar 2007.

[285] Rui Xu and D. Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645 –678, may 2005.

[286] Rui Xu and D. Wunsch. *Clustering*. IEEE/Wiley, 2009.

[287] Rui Xu and D.C. Wunsch. Clustering algorithms in biomedical research: A review. *Biomedical Engineering, IEEE Reviews in*, 3:120, 2010.

[288] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jörg Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings of the Fourteenth International Conference on Data Engineering*, ICDE '98, pages 324–331, Washington, DC, USA, 1998. IEEE Computer Society.

[289] Ronald R. Yager. Intelligent control of the hierarchical agglomerative clustering process. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 30(6):835–845, 2000.

[290] Nozomu Yanaihara, Natasha Caplen, Elise Bowman, Masahiro Seike, Kensuke Kumamoto, Ming Yi, Robert M Stephens, Aikou Okamoto, Jun Yokota, Tadao Tanaka, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell*, 9(3):189–198, 2006.

[291] Zhao Yanchang and Song Junde. GDILC: a grid-based density-isoline clustering algorithm. In *Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on*, volume 3, pages 140–145, 2001.

[292] Da Yang, Yan Sun, Limei Hu, Hong Zheng, Ping Ji, Chad V Pecot, Yanrui Zhao, Sheila Reynolds, Hanyin Cheng, Rajesha Rupaimoole, et al. Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell*, 23(2):186–199, 2013.

[293] Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making (IJITDM)*, 05(04):597–604, 2006.

[294] Eng J. Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon H. Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell*, 1(2):133–143, March 2002.

[295] Zhiwen Yu, Hau-San Wong, and Hongqiang Wang. Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, 23(21):2888–2896, 2007.

[296] Stefanos Zafeiriou and Nikolaos A. Laskaris. On the improvement of support vector techniques for clustering by means of whitening transform. *Signal Processing Letters, IEEE*, 15:198–201, 2008.

[297] Aidong Zhang. *Advanced analysis of gene expression microarray data*, volume 1. World Scientific Publishing Co Inc, 2006.

[298] Dao-Qiang Zhang and Song-Can Chen. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artificial Intelligence in Medicine*, 32(1):37–50, 2004.

[299] Chuan Zhou. *A Bayesian Model for Curve Clustering with Application to Gene Expression Data Analysis*. PhD thesis, University of Washington, 2003.

[300] Shangming Zhou and John Q. Gan. An unsupervised kernel based fuzzy c-means clustering algorithm with kernel normalisation. *International Journal of Computational Intelligence and Applications*, 04(04):355–373, 2004.

[301] H. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao. Gaussian mixture density modeling, decomposition, and applications. *IEEE Trans. Image Processing*, 5(9):1293–1302, September 1996.

[302] H-J Zimmermann. Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):317–332, 2010.