# Data Clustering Technologies In Cancer Subtyping

YAN YAN, BOBBY D. BRYANT, and FREDERIC C. HARRIS, JR., University of Nevada, Reno

Cancer subtyping remains a challenging task in microarray data analysis. The major goals of a successful cancer subtyping system are accuracy and reliability. Cluster analysis techniques have proven to be effective in this area. To facilitate further development in cancer subtyping based on microarray data, we provide a comprehensive review of the major cluster analysis algorithms from the clinical and computational domains that have been applied on microarray mRNA expression data and miRNA expression data for cancer subtyping, as well as other clustering algorithms with potential application in cancer subtyping.

CCS Concepts:•**Computing methodologies** → **Cluster analysis;** *Unsupervised learning;* Machine learning; Learning paradigms;•**Applied computing** → **Health informatics;** *Life and medical sciences;*

Additional Key Words and Phrases: algorithms, cancer subtype detection, data clustering, microarrays

**ACM Reference Format:**
Yan Yan, Bobby D. Bryant, and Frederic C. Harris, Jr. 2016. Data Clustering Technologies In Cancer Subtyping. *ACM Comput. Surv.* 1, 1, Article 1 (January 1111), 50 pages.
DOI: 0000001.0000001

## 1. INTRODUCTION

Clustering microarray gene expression data has been used to improve cancer subtyping [Alon et al. 1999; Golub et al. 1999; Slonim et al. 2000] 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 expressions) 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 [Slonim et al. 2000] and possibilities for more accurate and reliable results.

In this review, we focus on the application of clustering techniques. 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.

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 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 [Guha et al. 1998]. Clustering has a long history, tracing back to Aristotle, and has been studied extensively since the 18th century [Hansen and Jaumard 1997]. It has a wide range of applications in natural sciences, engineering, economics,

Author's addresses: Y. Yan, B. D. Bryant, and F. C. Harris, Jr., Department of Computer Science and Engineering, University of Nevada, Reno; emails: yyan66@yahoo.com, bdbryant@cse.unr.edu, and Fred.Harris@cse.unr.edu.

DOI: 0000001.0000001

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 [Hansen and Jaumard 1997].

Cluster analysis algorithms draw upon statistics, mathematics, and computer science [Hansen and Jaumard 1997]. Closely related fields are machine learning, pattern recognition, computer vision, image analysis, information retrieval, and bioinformatics. The k-means algorithm, first published in 1955 [Steinhaus 1956], 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 [Kuncheva and Hadjitodorov 2004], and discovering all cluster structures in a data set is impossible for any known clustering algorithm [Duda et al. 2001; Handl et al. 2005].

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 [Alon et al. 1999; Golub et al. 1999; Slonim et al. 2000], and the results are promising with improved accuracy over traditional methods [Slonim et al. 2000]. This kind of analysis was first employed in [Golub et al. 1999] and [Alizadeh et al. 2000]. 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 [D'haeseleer 2005].

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 [Alizadeh et al. 2000; Bhattacharjee et al. 2001; Bittner et al. 2000; Bredel et al. 2005; Chen et al. 2002; Chowdary et al. 2006; Dyrskjøt et al. 2002; Golub et al. 1999; Laiho et al. 2006; Lapointe et al. 2004; Liang et al. 2005; Risinger et al. 2003; Singh et al. 2002; Yeoh et al. 2002]. Novel algorithms specifically targeting gene expression data and taking its intrinsic characteristics into account were presented to improve the clustering results [Brunet et al. 2004; Liu et al. 2003; McLachlan et al. 2002].

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. 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 provides a historic background on cancer subtyping that includes very different subtyping techniques in the medical field. Section 3 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 4 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 5 examines challenges in microarray-data-based cancer subtyping. Finally, Section 6 concludes the review and points out future directions.

## 2. CANCER SUBTYPING TECHNIQUES

Many diseases, including cancers, have multiple subtypes. For example, leukemia has four main subtypes: acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML), chronic lymphocytic leukemia, and chronic myelogenous leukemia [NCI 2013a]. Each of these main subtypes can have several subsubtypes as well. For example, ALL has six subsubtypes [Society 2013]. Accurate subtyping can identify the most appropriate treatments to target specific disease subtypes, thus improving treatment results.

Given the tremendous complexity of various types and subtypes of cancers, it is believed that the single most important factor for surviving or managing cancers is and will be early detection and treatment [Xu et al. 2008]. Early cancer diagnosis requires accurately identifying the type as well as the subtype of a tumor. Tumors with similar appearances under microscope may have quite different origins and may therefore respond differently to the same treatment therapy. For example, DLBCL, the most common type of lymphoma in adults, can be cured by chemotherapy in only 35-40 percent of patients due to the existence of unknown subtypes that cannot be discriminated based only on their morphological parameters [Alizadeh et al. 2000].

The conventional approach to cancer therapy has been to provide treatment according to the organ or tissue in which the cancer originates. In the past two to three decades, however, the genetic events that lead to cancer have been dissected, and it has become clear that cancer develops as a result of multiple genetic defects and that individuals with the same type of cancer often have dissimilar genetic defects in their tumors. This finding explains why patients who seem to have similar cancer respond in a heterogeneous manner to anticancer agents and shows clearly the huge obstacle to providing effective treatments for cancer [Van't Veer and Bernards 2008]. New approaches to cancer therapy are shifting to a more personalized approach in which each patient is treated according to the specific genetic defects in the tumor.

Traditional cancer subtyping methods are largely dependent on the morphological appearance of tumors and parameters derived from clinical observations including histology, histopathology, and immunohistochemistry. The advent of DNA-microarray technology in the 1990s and the systematic analysis of the gene-expression patterns of tumor samples enabled researchers to investigate a new class of molecular diagnostic test for cancer [Van't Veer and Bernards 2008]. The gene expression profiles from particular microarray experiments have been recently used for cancer classification [Alon et al. 1999; Golub et al. 1999; Slonim et al. 2000]. This approach promises to give a better therapeutic measurement to cancer patients by diagnosing cancer types with improved accuracy [Slonim et al. 2000]. This section provides brief summaries of traditional subtyping methods, and focuses on the newer method of molecular diagnosis subtyping and applying clustering to molecular diagnosis subtyping.

### 2.1. Histology and Histopathology

Histology studies the microscopic anatomy of cells and tissues of plants and animals. It is commonly performed by sectioning and staining cells and tissues, then examining such cells and tissues with a light microscope or electron microscope [Martini 2012]. Marcello Malpighi (1628-1694), an Italian anatomist, conducted work that laid the foundation of histology. He used a rudimentary microscope to examine tissues and was the first to observe capillaries [DiDio 1994]. Marie François Xavier Bichat (1771-1802), a French anatomist, is generally regarded as the 'father of histology'. He suggested that body organs are composed of various tissues or membranes and introduced 21 such tissues [Roeckelein 1998]. In the early 19th century, histologist Camillo Golgi invented the tissue staining technique to enhance the contrast between cellular components [Prize 2013]. Histology is an essential tool in the fields of biology and medicine.

Histopathology studies the microscopic anatomy of diseased tissues in order to examine the manifestations of disease. Specifically, in clinical medicine, histopathology refers to the examination of a surgical specimen or biopsy by a pathologist after the specimen has been processed and placed onto glass slides [Allen 2013]. It is an important tool in anatomical pathology, and accurate diagnosis of cancer and other diseases usually requires histopathological examination of tissues.

### 2.2. Immunohistochemistry

Immunohistochemistry (IHC) refers to the process of identifying antigens (*e.g.* proteins) in cells of tissues. *'Immuno'* means antibodies, and *'histo'* means tissue. It is based on the principle that antibodies bind specifically to antigens in biological tissues. In an IHC process, antibodies are tagged with a visible label. IHC uses anatomical, immunological and biochemical techniques to visualize specific cellular components' distribution and localization within cells and tissues [Piercenet 2013].

As early as the 1930s, the principle of IHC was known [Duraiyan et al. 2012]. In 1942, Coons reported the first IHC study that identifies pneumococcal antigens in infected tissue using FITC-labeled antibodies [Coons et al. 1942]. Since then, improvements have been made in IHC, and IHC has become a routine and essential tool in laboratories for diagnosis and research purposes.

IHC can be used for disease diagnosis, drug development, and biological research. For example, physicians use IHC and specific tumor markers to diagnose a cancer as benign or malignant, as well as the tumor stage and grade, cell type and origin (site of the primary tumor) [Duraiyan et al. 2012].

### 2.3. Molecular Diagnosis

Conventional diagnosis of cancer type and subtype has been based on examination of the morphological appearance of stained tissue specimens under the light microscope, which is subjective and depends on pathologists' experiences. Diagnosis of cancer based on microarray data gives hope that cancer classification can be objective and highly accurate, which could provide clinicians enough information to choose the most appropriate treatment. The accuracy of morphological and immunohistochemistry diagnosis is insufficient in many diseases (including cancers) to identify diagnostic subtypes. Research has shown that a genomic method using gene expression in diseases can increase the clarity of previously obscure diagnostics. The molecular classification of diseases on the basis of gene expression can discover previously undetected but clinically significant subtypes of diseases [Alizadeh et al. 2000].

In 1949, Pauling et al. introduced the term 'molecular disease', which is based on their discovery that a single amino acid molecule change at the b-globin chain leads to the disease of sickle cell anemia [Pauling et al. 1949]. The foundation of molecular di-

agnostics was laid by scientists from various disciplines working on recombinant DNA technology in the 1970s [Chehab 1993]. In the 1980s, molecular diagnostics started to become popular with the discovery of the powerful molecular biology tool, Polymerase Chain Reaction [Bartlett and Stirling 2003]. In 1996, the application of microarray technology made it possible to simultaneously monitor thousands of genes in one experiment. Molecular diagnostics has been implemented not only in research laboratories but also in clinics on individual patients due to reduced complexities, costs, and time requirements involved in the process.

Molecular diagnosis of human diseases is nucleic acid based. It detects various pathogenic mutations in nucleic acid (*i.e.* DNA and/or RNA samples) in order to facilitate detection, diagnosis, subtyping (or subclassification), prognosis of diseases, and monitoring response to therapy [Chehab 1993]. Molecular diagnostics uses technologies such as mass spectrometry and gene chips from the fields of genomics and proteomics to classify cancers and other diseases. Molecular diagnostics studies how the genes and proteins interact in a cell and discovers their activity patterns by examining the expression patterns of genes. For example, in different types of cancerous or precancerous cells, molecular diagnostics uncovers differences in gene expression patterns that can be used to improve cancer diagnosis [NCI 2013b].

Ideally, measurement of gene expression is done on the final gene product, the protein for many genes. However it is often easier to use one of the precursors, typically mRNA (messenger RNA), to infer gene expression level [Xu 2011]. There are different technologies to measure gene expression including reporter gene [Koo et al. 2007], northern blotting [Alberts et al. 2007; Kevil et al. 1997], western blotting [Song et al. 2009], RT-qPCR [Taylor et al. 2010], SAGE [Velculescu et al. 1995], and DNA microarray [Schena et al. 1995].

DNA Microarrays, also called 'gene chips', are an efficient way to measure gene expression, allowing researchers to obtain the expression of thousands of genes at once. When active genes assemble proteins, they produce mRNA molecules. The mRNA molecules' strands are then copied to cDNAs (complementary DNAs) using the DNA base-pairing rules. The resulting cDNAs are then attached to fluorescent dye. When the fluorescent cDNAs are put into a DNA microarray, each cDNA molecule will bind to the microarray spot containing pieces of its specific matching gene by base-pairing. This way, each microarray spot corresponds to a gene which was actively being transcribed into mRNA in the original cell. The intensity and color of the fluorescence of each microarray spot are measured and analyzed by a computer program to make precise measurements of genes expressed at different levels. Further information about microarray technology can be found in [Müller and Nicolau 2005].

Microarrays and molecular diagnosis have been used in cancer research and have produced very promising results. For example, researchers used microarrays to compare gene expression patterns in normal vs. cancerous stomach tissues and found that the PLA2G2A gene is expressed in the cancerous stomach tissue only. The researchers also found that stomach cancer patients with high expression levels of PLA2G2A had better 5-year survival prognosis than those with low expression levels [NCI 2013b]. Researchers also use microarrays to detect the differences in gene expression patterns within the same cancer type. In many cases, researchers have found that a single type of cancer based on microscope observations contains several subtypes based on their different gene expression patterns. For example, researchers have identified several subtypes of leukemia and four subtypes of lung adenocarcinoma, the most common type of lung cancer.

Besides being used for screening and diagnosing diseases, molecular diagnostics has many other uses. For example, expression patterns can provide information for new treatment design, treatment effectiveness evaluation, and patient response prediction.

## 3. 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 [Jain 2010]. Cluster analysis is defined as 'a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics' and its first known use was in 1948 (Merriam-Webster Online Dictionary, 2013). The clustering algorithm was first developed by biologists in numerical taxonomy study in 1963 before being utilized by statisticians [Ilango et al. 2011]. Clustering is used for class discovery, *i.e.* exploration or discovery of the underlying patterns of a dataset by separating the dataset into groups, with little or no prior knowledge [Everitt et al. 2001; Jain and Dubes 1988; Xu and Wunsch 2005; 2009]. Clustering is also used for natural classification, *i.e.* identifying the degree of similarity among organisms, and compression, *i.e.* organizing and summarizing data using cluster prototypes [Jain 2010]. Clustering has become increasingly popular as 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 [Linden 2009; Xu and Wunsch 2010].

There is no universally agreed upon definition of clusters [Everitt et al. 2001]. 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 [Jain 2010]. 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 [Ilango et al. 2011]. 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.

Since the process of clustering is subjective, judging the relative efficacy of clustering methods is difficult [Baraldi and Alpaydin 2002; Jain et al. 1999]. Cluster validity is used to assess clustering results and can be classified into three categories: a) Internal validities formulate quality as a function of the given data set [Hu and Yoo 2004]. Examples include Dunn's Validity Index, Silhouette Value, Hubert Gamma Statistic, Entropy, Xie-Beni, Normalized Mutual Information. b) External validities assess quality by additional external information such as category labels [Hu and Yoo 2004]. Examples include Jaccard Index, Rand Index, Adjusted Rand Index, Variation of Information, Kappa Statistic, CA. c) Relative validities evaluate a clustering result by comparing it to results from other clustering methods.

The procedure of cluster analysis includes four steps [Xu and Wunsch 2005]: a) Step one is feature selection or extraction. Feature selection selects a subset of all features, and feature extraction generates novel features from the original ones by using some transformations [Bishop 1995; Jain et al. 1999; Jain et al. 2000; Xu and Wunsch 2005]. b) Step two is clustering algorithm design or selection. Since clustering algorithms group objects based on some proximity measure, this step usually includes choosing an appropriate proximity measure and construction of a clustering criterion function, creating an optimization problem that has been well studied in the literature. c) Step three is cluster validation. This step calculates a confidence level for the clustering results. d) Step four is results interpretation. This step provides meaningful insights from the data.

There is no single clustering algorithm that performs best across all problems or data sets [Kleinberg 2002; Xu and Wunsch 2005]. Therefore, it is important to study the characteristics of the problem and use an appropriate clustering strategy [Xu and Wunsch 2005].

Properties to be considered in choosing a clustering algorithm include [Berkhin 2002]: 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.

Despite many examples of successful applications of cluster analysis, there still remain many challenges due to the existence of many inherent uncertain factors [Xu and Wunsch 2005]. The following fundamental challenges in clustering [Jain and Dubes 1988; Jain 2010] are relevant even today [Jain 2010]: 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.

Some recent trends in clustering include [Jain 2010]: semi-supervised clustering utilizing external or side information; interactive clustering, where a user can specify or change program parameters based on domain knowledge or results from previous clustering iterations; clustering ensembles, where the partitions resulting from different algorithms (or the same algorithm with different parameters) are combined; multi-objective clustering, where the clustering algorithm optimizes multiple specific objectives; large-scale clustering, which handles very large databases; multi-way clustering, which extends the bi-clustering framework and simultaneously clusters heterogeneous components of the data objects [Bekkerman et al. 2005]; and heterogeneous data clustering for data comprising multiple types, such as rank data, dynamic data, graph data, and relational data [Ilango et al. 2011].

Clustering techniques can be organized into categories. Different criteria may result in different categories of clustering algorithms [Xu and Wunsch 2005]. Furthermore, categorization of clustering algorithms is not straightforward or canonical, and categories can overlap [Berkhin 2002]. For convenience, in this review we use the following taxonomy, which is also widely used in the literature: hierarchical clustering (Section 3.1), partitioning clustering (Section 3.2), graph-based Clustering (Section 3.3), distribution-based clustering (Section 3.4), density-based clustering (Section 3.5), grid-based clustering (Section 3.6), clustering big data (Section 3.7), clustering high dimensional data (Section 3.8), and other clustering techniques (Section 3.9).

### 3.1. Hierarchical Clustering

Hierarchical clustering algorithms organize a data set into a hierarchical structure according to a similarity measure [Xu and Wunsch 2005]. It is based on the belief that nearby objects are more related than objects that are farther away [Nangia 2012]. 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 [Berkhin 2002].

Divisive clustering is very expensive in computation [Everitt et al. 2001] and is not commonly used in practice [Xu and Wunsch 2005]. We focus on the agglomerative

clustering first and then mention two divisive clustering algorithms named MONA and DIANA [Kaufman and Rousseeuw 2009; Xu and Wunsch 2005].

There are many agglomerative hierarchical clustering algorithms based on different linkage criterion. The single linkage method or nearest neighbor method [Gower and Ross 1969; Jain and Dubes 1988; Sibson 1973; Sneath and Sokal 1973; Sneath 1957; Xu and Wunsch 2005] uses the distance between two closest objects in different clusters, and the shortest distance determines the merge of two clusters. The complete linkage method or farthest neighbor method [Defays 1977; King 1967; Sorensen 1948; Xu and Wunsch 2005] uses the distance between two farthest objects in different clusters, and the shortest distance determines the merge of two clusters. These two methods are the simplest and most popular [Xu and Wunsch 2005]. Average linkage methods include UPGMA (Unweighted Pair-Group Method using Arithmetic averages), WPGMA (Weighted Pair-Group Method using Arithmetic averages), UPGMC (Unweighted Pair Group Method using Centroids), and WPGMC (Weighted Pair Group Method using Centroids). UPGMA and UPGMC use a simple average, while WPGMA and WPGMC use a weighted average where the weight is the inverse of cluster size. UPGMA [D'Andrade 1978; Everitt 1980; Jain and Dubes 1988; Sneath and Sokal 1973; Sokal and Michener 1958] uses average distance between two objects in different clusters, and the shortest average distance determines the merge of two clusters. WPGMA or weighted average linkage method [Murtagh 1983] uses weighted average distance between two objects in different clusters, and the shortest average distance determines the merge of two clusters. UPGMC or centroid linkage method [Sneath and Sokal 1973] uses Euclidean distance between unweighted centroids (calculated by arithmetic mean) of different clusters, and the shortest distance determines the merge of two clusters. WPGMC or median linkage method [Sneath and Sokal 1973] uses Euclidean distance between weighted centroids of different clusters, and the shortest distance determines the merge of two clusters. Minimum-variance method or Ward's method [Ward Jr. 1963] 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 [Wildi 2010]. Further readings about these methods include [Everitt et al. 2001; Xu and Wunsch 2005; Yager 2000].

What follows are examples of divisive hierarchical clustering algorithms. DIANA [Kaufman and Rousseeuw 2009] (DIvisive ANAlysis Clustering) selects in each dividing step the cluster with the largest diameter and divides it into two new clusters. MONA [Kaufman and Rousseeuw 2009] (MONothetic Analysis Clustering of Binary Variables) divides clusters based a single well-chosen variable (or feature), whereas most other hierarchical methods use all variables (or features).

Advantages of hierarchical clustering are a) Good visualization with dendrogram representation [Xu and Wunsch 2005; 2010; Jain and Dubes 1988; Theodoridis and Koutroumbas 2006], b) Very informative descriptions with dendrogram representation [Jain and Dubes 1988; Theodoridis and Koutroumbas 2006; Xu and Wunsch 2005; 2010], and c) Flexibility regarding the number of clusters, since the clustering results can be obtained by cutting the dendrogram at different levels.

Disadvantages of hierarchical clustering are [Xu and Wunsch 2005; 2010]: a) Lacking of robustness and sensitivity to noise and outliers. b) High computational complexity, which limit their application on large scale data. c) Tendency to form clusters with spherical shapes instead of natural shapes. d) Prone to reversal phenomenon [Okabe and Sugihara 2012].

BIRCH [Zhang et al. 1996] (Balanced Iterative Reducing and Clustering using Hierarchies) clusters incoming data objects incrementally and dynamically. It first builds a CF (Clustering Feature) tree dynamically as new data objects are inserted and then applies an agglomerative hierarchical clustering algorithm to the nodes represented

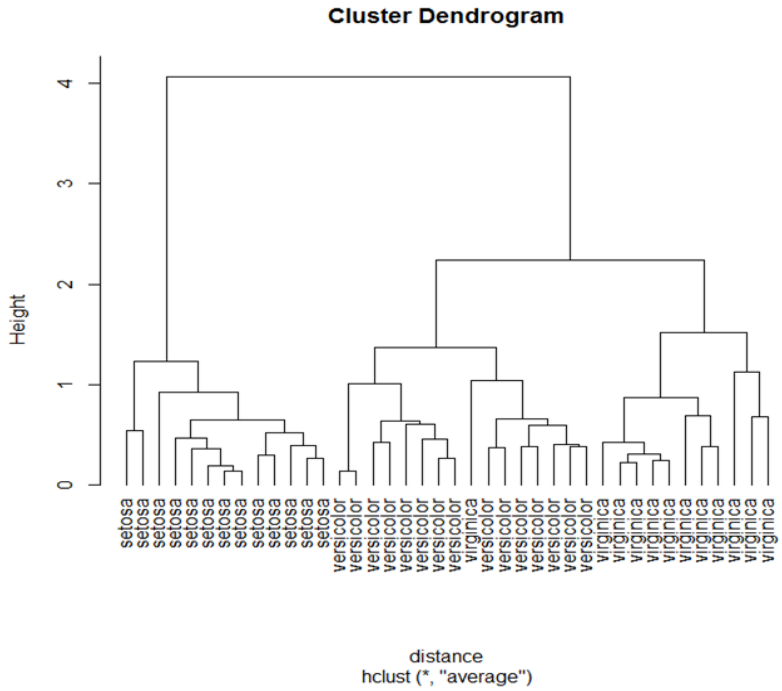


Fig. 1. Hierarchical Clustering [Ho 2012]

by their CF vectors. After obtaining a centroid for each cluster, it assigns each data object to its nearest centroid. CURE [Guha et al. 1998] (Clustering Using REpresentatives) uses a number of representative data points in a cluster to evaluate the distance between clusters. Closest cluster pair are merged at each step of its hierarchical clustering process. ROCK [Guha et al. 2000] (RObust Clustering using linKs) uses links and not distances when merging clusters for boolean and categorical data. DISMEA [Spath 1980] uses the k-means algorithm to divide a cluster into two clusters. The Edwards and Cavalli-Sforza Method [Edwards and Cavalli-Sforza 1965] divides all available clusters at each step. Minimum Spanning Tree-based clustering algorithms [Eldershaw and Hegland 1997; Päivinen 2005; Zahn 1971] construct an MST (Minimum Spanning Tree) [Kruskal 1956; Nešetřil et al. 2001; Prim 1957] from a data set and produce a group of clusters by removing selected edges.

Figure 1 shows an example of hierarchical clustering.

### 3.2. 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.

k-means clustering [Berkhin 2002; Forgy 1965; Hartigan 1975; Hartigan and Wong 1979; MacQueen 1967; Steinhaus 1956; Xu and Wunsch 2005; 2010] is very simple, but one of the best known and popular clustering algorithms. There are many variations of the basic k-means clustering. Classic k-means reassigns data objects based on

optimization of the objective function. If a reassigning has positive effect, the data object is reassigned and the cluster centers are updated. ISODATA [Ball and Hall 1965] (Iterative Self-Organizing Data Analysis Technique) splits and merges intermediate clusters based on a user-defined threshold and iterates until the threshold is reached. FORGY [Forgy 1965] reassigns objects to nearest centroids and recomputes centroids. It iterates until a stopping criterion is achieved. Fuzzy c-means [Bezdek et al. 1981; Dunn 1973] 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) [Eschrich et al. 2003] replace similar data objects with their centroid before clustering.

Variations of k-medoid [Kaufman and Rousseeuw 2009] 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) [Kaufman and Rousseeuw 2009] 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) [Ng and Han 1994] 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 are 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 are 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).

Other developments are as follows. Bisecting k-means [Steinbach et al. 2000] recursively partitions a cluster into two. KD-trees k-means [Pelleg and Moore 1999] uses the KD-Tree data structure to speed up the assignment of data objects to their closest cluster by reducing the number of nearest-neighbor queries in the traditional algorithm. Scaling k-means [Bradley et al. 1998] retains important data objects and summarizes or discards other objects. Centroids of the resulting data set are then used on the whole data set. X-means [Pelleg and Moore 2000] finds the number of clusters K automatically by optimizing a criterion function such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). Kernel k-means [Schölkopf et al. 1998] enhances k-means by using a kernel function that nonlinearly maps the original feature space to a higher dimensional one, where clusters are more separable. Weighted kernel k-means [Dhillon et al. 2004] further extends kernel k-means by assigning a weight for each cluster. The weight is defined as the reciprocal of the number of data objects in the cluster. GA k-means [Babu and Murty 1993] applies a genetic algorithm to improve cluster centers initialization for k-means. Simulated annealing [Al-Sultan and Selim 1993; Huntley and Brown 1991; Klein and Dubes 1989; Selim and Alsultan 1991] uses simulated annealing optimization to avoid local optima and find the global minimum solution. Soft assignment [Zhang 2001] assigns data objects to different clusters with appropriate weights to improve the optimization process. It uses Harmonic Averages of the distances from the data object to all the centers. Mahalanobis distance [Mao and Jain 1996] is used to detect clusters with hyperellipsoidal shapes. Maximum of intracluster variances [Gonzalez 1985] can be used as the objective function instead of the sum to obtain good clustering results. k-prototypes [Huang 1998] incorporates cate-

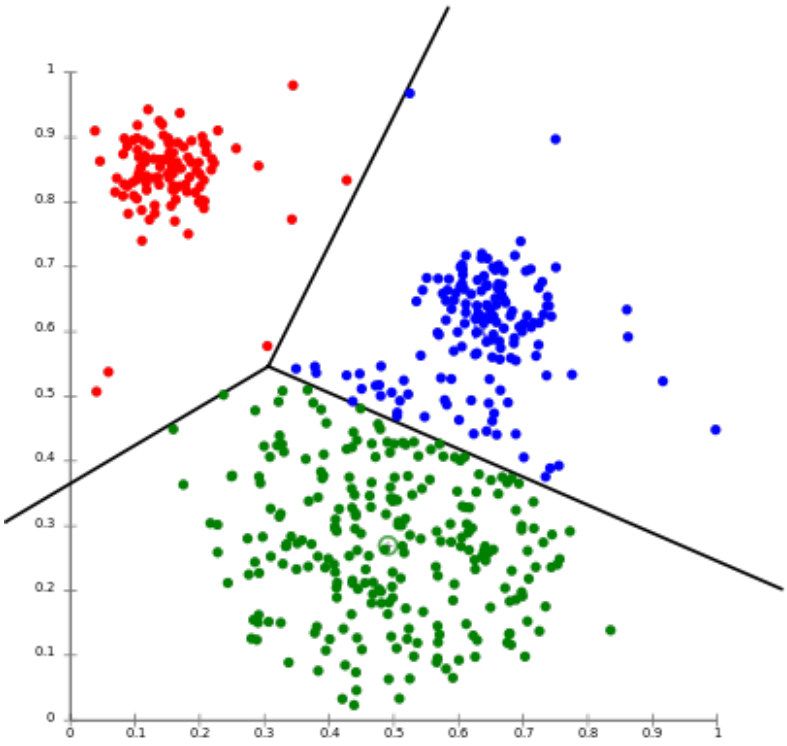


Fig. 2. Partitioning Clustering [Chire 2011c]

gorical data as a generalization approach. Accelerated k-means by triangle inequality [Elkan 2003] avoids unnecessary distance calculations by using the triangle inequality and keeping track of lower and upper bounds for distances between data objects and cluster centers. k-means++ [Arthur and Vassilvitskii 2007] improves the speed and the accuracy of k-means by using a simple randomized seeding technique.

Figure 2 shows an example of partitioning clustering.

### 3.3. 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 [Chen 2010]. The edges in the same subgraph/subhypergraph should have high weights, and the edges between different subgraphs/subhypergraph should have low weights [Chen 2010]. It is also called spectral clustering [Jain 2010].

Representative algorithms are as follows. Chameleon [Karypis et al. 1999] 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 Categor-

ical Data Using Summaries) [Ganti et al. 1999a] 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 [Gan 2011]. A Dynamic System-based Approach or STIRR (Sieving Through Iterated Relational Reinforcement) [Gibson et al. 2000] 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) [Guha et al. 2000] 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 [Gan 2011].

The advantages of graph-based clustering are [Chen 2010]: 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. 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 [Chen 2010].

Other developments are as follows. The Ratio Cut algorithm [Hagen and Kahng 1992] adopts a cluster size constraint, which is the number of data points in a cluster. The Normalized Cut (NCut) algorithm [Shi and Malik 2000] is an approximate graph-cut based clustering algorithm with a cluster size constraint, which is the volume of the cluster or sum of edge weights within a cluster. It also has a multiclass version [Yu and Shi 2003]. The MNCut (Modified Normalized Cut) algorithm [Meila and Shi 2001] gives a new interpretation to the NCut algorithm in the framework of a Markov Random Walk. Ng's method [Ng et al. 2001] derives a new data representation from normalized eigenvectors of a kernel matrix simultaneously and in a particular manner. Laplacian Eigenmap [Belkin and Niyogi 2003] uses the eigenvectors of the graph Laplacian to represent data. Pairwise Data Clustering by Deterministic Annealing [Hofmann and Buhmann 1997] uses proximity measures between the data objects to represent data. Dominant Sets Pairwise Clustering [Pavan and Pelillo 2007] relates clusters to maximal dominant sets [Motzkin and Straus 1965] in pair-wise clustering. Fast approximate spectral clustering [Yan et al. 2009] applies a distortion-minimizing local transformation to the data to speed up conventional spectral clustering. Active spectral clustering [Wang and Davidson 2010] follows the concept of constrained clustering and uses pairwise relations. Its constraints are specified in a incremental manner. Locally-scaled spectral clustering using empty region graphs [Correa and Lindstrom 2012] employs $\beta$-skeleton (a subset of empty region graphs) and non-linear diffusion to define a locally adapted affinity matrix which defines the similarity of two data objects.

Figure 3 shows an example of graph-based clustering.

### 3.4. 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 [Gan et al. 2007; McLachlan and Basford 1988]. 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 [Gan et al. 2007]. 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 [Gan et al. 2007].

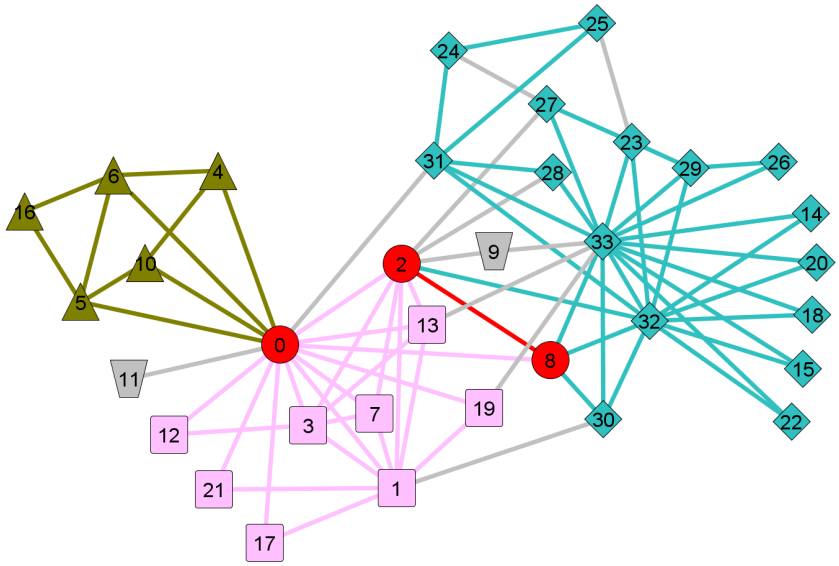


Fig. 3. Graph-based Clustering [Evans 2014]

Distribution-based clustering has a long history. Early works include [Binder 1978; Day 1969; Scott and Symons 1971; Wolfe 1970]. A survey of cluster analysis in a probabilistic and inferential framework is presented in [Bock 1996].

Representative algorithms are as follows. The EM (Expectation-Maximization) clustering algorithm [Dempster et al. 1977] 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)[Barbara et al. 2002] uses entropy to cluster categorical data. It consists of data sampling and incremental assignment. STUCCO (Search and Testing for Understandable Consistent Contrasts) [Bay and Pazzani 1999] uses tree searching and significant contrast-sets to find clusters. GMDD (Gaussian Mixture Density Decomposition) [Zhuang et al. 1996] uses a recursive approach and identifies each Gaussian component in the mixture successively. Autoclass [Cheeseman and Stutz 1996] is based on the classic distribution-based approach and uses a Bayesian method to determine the optimal clusters. P-AutoClass [Pizzuti and Talia 2003] is a parallel version of Autoclass and can be used on large data sets.

The advantages of distribution-based clustering are as follows: a) It can be modified to handle complex data [Berkhin 2002], b) It has a solid theoretical foundation, c) Its results are easily interpretable [Berkhin 2002], 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 [Berkhin 2002], f) It is good for online learning since the intermediate mixture model can be used to cluster objects [Berkhin 2002], g) the Mixture model can be naturally generalized to cluster heterogeneous data [Berkhin 2002].

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).

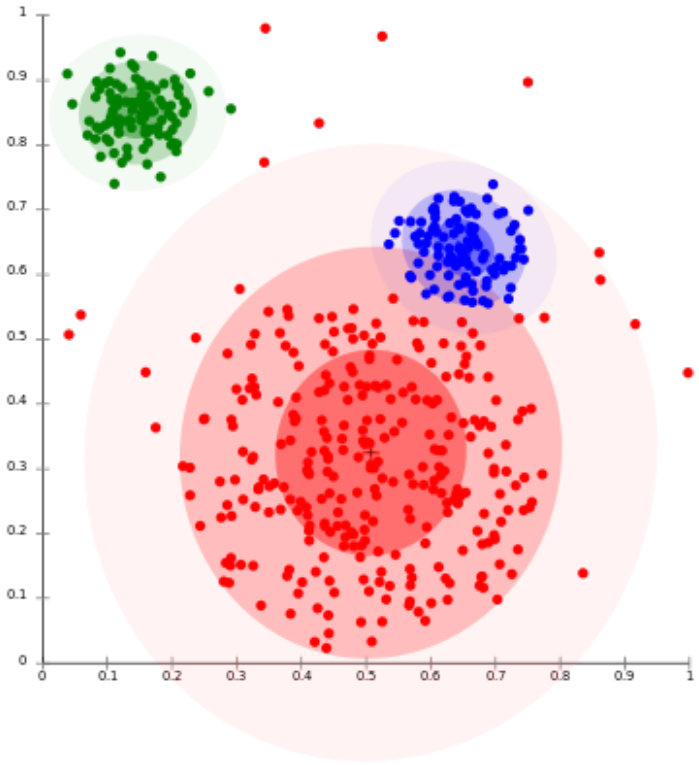


Fig. 4. Distribution-based Clustering [Chire 2011b]

Other developments are as follows. Latent Dirichlet Allocation (LDA) [Blei et al. 2003] uses a hierarchical Bayesian model that has three levels. Each data object is modeled as a finite mixture over an underlying set of groups (or clusters) of objects. Each group (or cluster) is modeled as an infinite mixture over a set of group (or cluster) probabilities. Pachinko Allocation Model (PAM) [Li and McCallum 2006] uses a Directed Acyclic Graph (DAG) to model cluster correlations. The leaves of the DAG represent data objects, and the interior nodes represent correlations. Undirected graphical model for data clustering [Welling 2005] is based on exponential family distributions and the semantics of undirected graphical models. It uses the technique of minimizing contrastive divergence to speed up the process. Robust cluster analysis via mixture models method [McLachlan et al. 2006] uses the mixtures of multivariate t distributions approach to the clustering. It also uses the t distribution to cluster high-dimensional data via mixtures of factor analyzers. Online learning for LDA method [Hoffman et al. 2010] is an online Variational Bayes (VB) algorithm for LDA. It uses natural gradient step in online stochastic optimization, which converges to a local optimum of the VB objective function.

Figure 4 shows an example of distribution-based clustering.

### 3.5. 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 [Gan et al. 2007]. 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 [Berkhin 2002]: 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) [Ester et al. 1996] 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) [Sander et al. 1998] 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) [Ankerst et al. 1999] 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) [Xu et al. 1998] 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.

Representative algorithms for the density function approach are as follows. DENCLUE (DENsity-based CLUstEring) [Hinneburg et al. 1998] calculates the impact of each data object within its neighborhood (*i.e.* influence function) and determines clusters mathematically by identifying local maxima of the overall density function (*i.e.* density-attractors).

The advantages of density-based clustering are as follows: a) They can find clusters of arbitrary shapes, in contrast to many other methods. b) Time complexity is low (linear or $\mathrm{O}(n)$). c) It is deterministic for core and noise points (but not for border points), therefore there is no need to run it multiple times. d) It can handle noise well [Gan et al. 2007]. e) The number of clusters is not required, since it finds clusters and the number of clusters automatically [Gan et al. 2007]. f) Results are independent of data ordering [Berkhin 2002]. g) There are no limitations on the dimension or attribute types [Berkhin 2002].

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 [Jain 2010]. d) Most density-based clustering algorithms were developed for spatial data [Gan et al. 2007].

Other developments are as follows. BRIDGE [Dash et al. 2001] integrates the k-means algorithm and the DBSCAN algorithm. k-means is first performed, and then DBSCAN is used on each partition. Finally, results are improved by removing the noise found by DBSCAN. JarvisPatrick algorithm [Frank and Todeschini 1994] partitions the data set into clusters based on the number of shared nearest neighbors. It first identifies the k nearest neighbors of each data object and then merges two data objects at a time. C-DBSCAN (Constrained-DBSCAN) [Ruiz et al. 2007] enhances the DB-

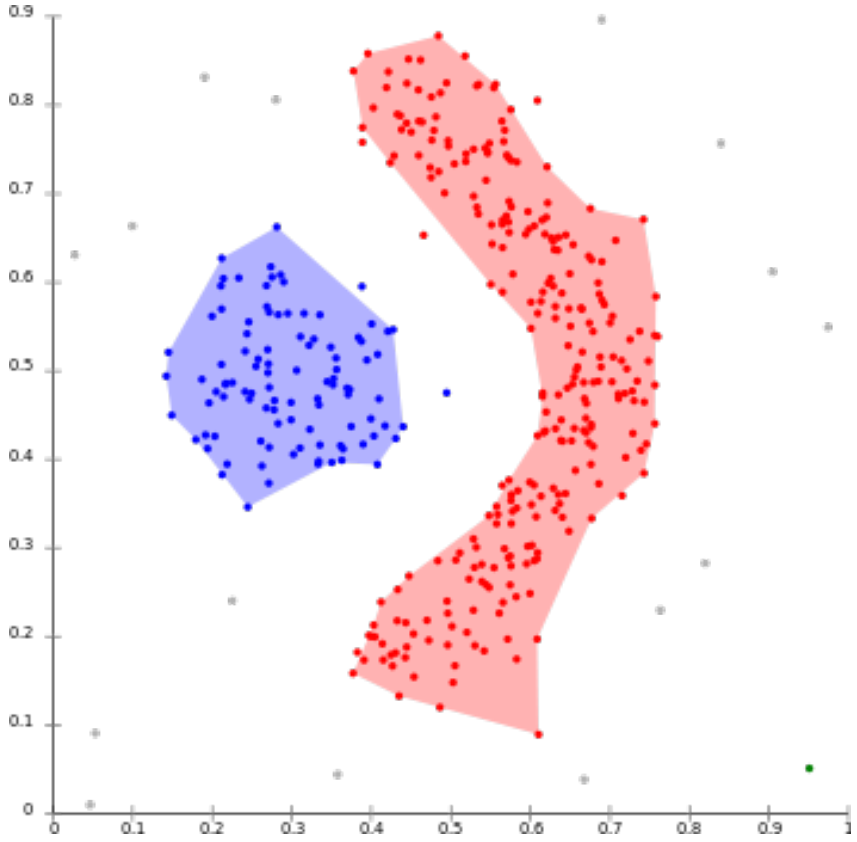


Fig. 5. Density-based Clustering [Chire 2011a]

SCAN algorithm with pairwise constraints. SCAN (Structural Clustering Algorithm for Networks) [Xu et al. 2007] can detect hubs and outliers, in addition to clusters in networks (or graphs). It uses a structural similarity measure to cluster vertices.

Figure 5 shows an example of density-based clustering.

### 3.6. Grid-based Clustering

Grid-based clustering operates on space partitioning instead of data partitioning to produce clusters [Berkhin 2002]. 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 [Berkhin 2002; Schikuta and Erhart 1997] 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) [Wang et al. 1997] 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+ [Berkhin 2002; Wang et al. 1999] uses a similar hierarchical cell structure as STING and introduces an active spatial data mining approach. OptiGrid (Optimal Grid) [Hinneburg and Keim 1999] 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) [Schikuta 1996] 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) [Yanchang and Junde 2001] 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) [Sheikholeslami et al. 1998] 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) [Barbara and Chen 2000] 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) [Berkhin 2002; Gan et al. 2007]. b) It handles noise well [Berkhin 2002]. c) It is independent of data ordering [Berkhin 2002]. d) It can handle attributes of different types [Berkhin 2002]. e) It can be used as an intermediate step in many other algorithms such as CLIQUE and MAFIA [Berkhin 2002].

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) [Gan et al. 2007]. b) Some grid-based clustering algorithms (*e.g.* STING, WaveCluster) are not good at high dimensional data [Gan et al. 2007].

Other developments are as follows. AMR (Adaptive Mesh Refinement clustering) [Liao et al. 2004] creates grids at multiple resolutions where higher resolution grids are applied to the localized denser regions. O-Cluster (Orthogonal partitioning CLUSTERing) [Milenova and Campos 2002] is a variant of OptiGrid. It creates a hierarchical grid-based structure by making axis-parallel (orthogonal) partitions on the input data. It operates recursively, and the final irregular grid frames the data into clusters. CBF (Cell-Based Filtering) [Chang and Jin 2002] splits each dimension into a set of partitions using a filtering-based index. It then creates cells based on the overlapping regions of the partitions. PGMCLU (Parallel Grid-based CLUstering algorithm for Multi-density datasets) [Xiaoyun et al. 2009] consists of parallel data partitioning, local clustering, and merging local clusters. It introduces a new measure called grid compactness for the degree of tightness between data objects within the grid, and the notion of grid feature for summarizing the information about a grid.

Figure 6 shows an example of grid-based clustering.

### 3.7. Clustering Big Data

Big data clustering refers to clustering on millions of data objects [Jain 2010]. These algorithms need to have good scalability and process big data within reasonable computing time and memory space [Berkhin 2002]. A high computational complexity would dramatically limit an algorithm's application to big data. The strategies used for big data clustering can be categorized into sampling, data summarization, distributed computing, and incremental learning.

*3.7.1. Sampling:.* Sampling methods select a sample of the original large data set and perform clustering over the sample data. Old-fashioned sampling methods may or may not use rigorous statistical reasoning [Berkhin 2002]. Newer sampling methods use special uniform checks to control their adequacy [Berkhin 2002]. Advantages are that it is simple to implement and can screen out most outliers. However, small clusters may be missed.

Examples are as follows. CURE (Clustering using REpresentatives) [Guha et al. 1998] and ROCK (RObust Clustering using linKs) [Guha et al. 2000] were covered in Section 3.1. CLARA (Clustering LARge Applications) [Kaufman and Rousseeuw 1987]

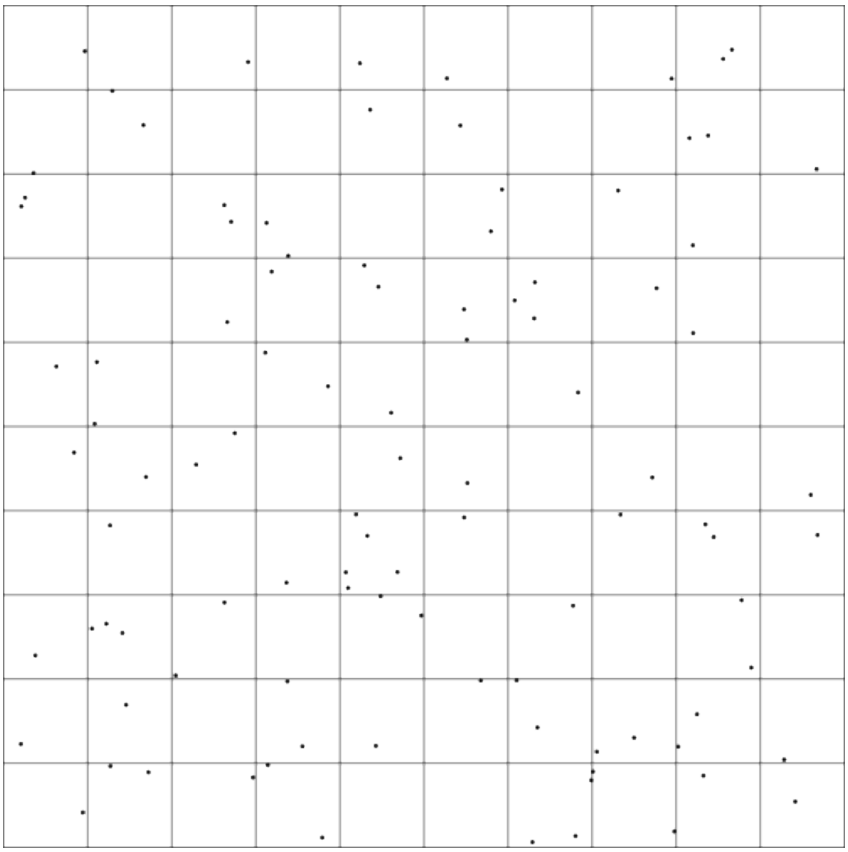

Fig. 6. Grid-based Clustering [Yolkowski 2014]

draws several samples from the data set, runs PAM on each of them, and selects the best result. CLARANS (Clustering Large Applications based on RANdomized Search) [Ng and Han 1994] starts with a new randomly-selected node (a set of k potential medoids) in the graph in search of the local optimum. It repeats if a local optimum is found.

*3.7.2. Data Summarization:.* Data summarization methods calculate data summary statistics and perform clustering on the summaries instead of the original data. The advantage is that the requirement for the storage of and frequent operations on the large amount of data are greatly reduced, saving both computational time and storage space. The disadvantage is reduced cluster quality.

Examples are as follows. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) was covered in Section 3.1. BUBBLE [Ganti et al. 1999b] instantiates generalized BIRCH for data in a distance space. BUBBLE-FM (BUBBLE-FastMap) [Ganti et al. 1999b] improves upon BUBBLE by reducing the computation time using FastMap [Faloutsos and Lin 1995]. EMADS (EM Algorithm for Data Summaries) [Jin 2002] directly generates a Gaussian mixture model from simplified data summaries. bEMADS (BIRCH's EMADS) [Jin 2002] uses data summarization procedures in the BIRCH algorithm.

*3.7.3. Distributed Computing:.* Distributed computing methods divide a large data set into smaller data sets and perform clustering on each smaller data set. The advantage is that clusterings on each smaller data set can be done in parallel to reduce the overall computation time [Jain 2010]. The disadvantage is the overhead and complexities due to the dividing and combining steps.

Examples are as follows. Parallel k-means [Dhillon and Modha 1999] is a parallel implementation of the k-means clustering algorithm. DBDC (Density Based Distributed Clustering) [Januzaj et al. 2004] clusters distributed data locally and extracts suitable representatives from these local clusters to send to a global site where the complete clusters are restored based on the local representatives. It uses a density-based clustering algorithm for both local and global clustering. Parallel spectral clustering in distributed systems [Chen et al. 2011] makes the dense similarity matrix sparse by retaining nearest neighbors using a parallel approach.

*3.7.4. Incremental Learning:.* Incremental learning methods process one data object at a time and may discard it. They require only one single pass over all data objects, in contrast to most clustering methods that require multiple passes over data objects before identifying the cluster centers [Jain 2010]. Advantages are: improved clustering efficiency in terms of data storage and processing time (they can admit new data objects without learning from scratch [Xu and Wunsch 2010]); handling outliers well [Berkhin 2002]; resumable processing which makes it very suitable for dynamic big data sets [Berkhin 2002]. Disadvantages are that results depend on data order and may not be stable [Carpenter and Grossberg 1987b; Moore 1989; Xu and Wunsch 2010], and can result in lower quality clusters [Berkhin 2002].

Examples are as follows. DIGNET [Thomopoulos et al. 1995; Wann and Thomopoulos 1997] moves cluster centers toward a new data point with each new addition. Hartigan's leader algorithm [Hartigan 1975] uses a distance/similarity threshold to decide if a data point should be added to the cluster or used for a new cluster center. ART (Adaptive Resonance Theory) family [Carpenter and Grossberg 1987a; Xu and Wunsch 2010] simulates neural circuits that are believed to trigger fast learning. It includes a large family of neural network variants such as ART1 [Carpenter and Grossberg 1987b], ART2 [Carpenter and Grossberg 1987a], Gaussian ART [Williamson 1996], Bayesian ART [Vigdor and Lerner 2007], Ellipsoid ART [Anagnostopoulos and Georgiopoulos 2001], ART tree [Caudell et al. 1991; Wunsch et al. 1993], ARTMAP [Carpenter et al. 1991], Q-learning ART [Brannon et al. 2009], Fuzzy ART [Carpenter et al. 1992]. Charikar's incremental clustering [Charikar et al. 1997] maintains a clustering of the data objects so that the maximum cluster diameter is minimized as new data objects are added. GenIC (Generalized Incremental algorithm for Clustering) [Gupta and Grossman 2004] divides the data stream into chunks or windows, updating each cluster center with each new data object addition and merging clusters at the end of a window of data. Cobweb [Fisher 1987] is an incremental system for hierarchical clustering, which enables bi-directional hill-climbing search through the space of hierarchical schemes.

### 3.8. 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 [Chakrabarti 2003]). Clustering high dimensional data is tremendously difficult. One problem is that increased irrelevant features eliminate the likelihood of clustering tendency [Berkhin 2002]. 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) [Berkhin 2002]. Performing feature selection before applying clustering can improve the first problem. Principal Component Analysis (PCA) [Pearson 1901] 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.

*3.8.1. 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 [Avogadri and Valentini 2009].

Examples are as follows. PreDeCon [Böhm et al. 2004] finds subsets of feature vectors that have low variance along subsets of attributes. PROCLUS [Aggarwal et al. 1999] 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 [Boutsidis et al. 2010] implements a dimensionality reduction technique for k-means clustering based on random projections.

*3.8.2. Subspace Clustering:.* Subspace clustering algorithms identify clusters in appropriate subspaces of the original data space.

Examples are as follows. CLIQUE (CLustering In QUEst) [Agrawal et al. 1998] partitions the data space into units and then finds the maximum sets of connected dense units. SUBCLU (density-connected Subspace Clustering) [Kröger et al. 2004] adopts the notion of density-connectivity introduced in DBSCAN (Section 3.5) and uses the monotonicity of density-connectivity to prune subspaces. CACTUS (Clustering Categorical Data Using Summaries) is covered in Section 3.3. ENCLUS (ENtropy-based CLUStering) [Cheng et al. 1999] 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) [Goil et al. 1999] uses adaptive grids in each dimension and then merges them to find clusters in higher dimensions. OptiGrid (Optimal Grid) is covered in Section 3.6. MrCC (Multi-resolution Correlation Cluster detection) [Cordeiro et al. 2010] 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.

Figure 7 shows an example of subspace clustering.

*3.8.3. 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) [Procopiuc et al. 2002] uses a global density threshold to compute an approximation of an optimal projective cluster. FIRES (FIlter REfinement Subspace clustering) [Kriegel et al. 2005] 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) [Moise et al. 2006; 2008] 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.

*3.8.4. Bi-clustering:.* Bi-clustering is also called bi-dimensional clustering [Cheng and Church 2000], 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 un-

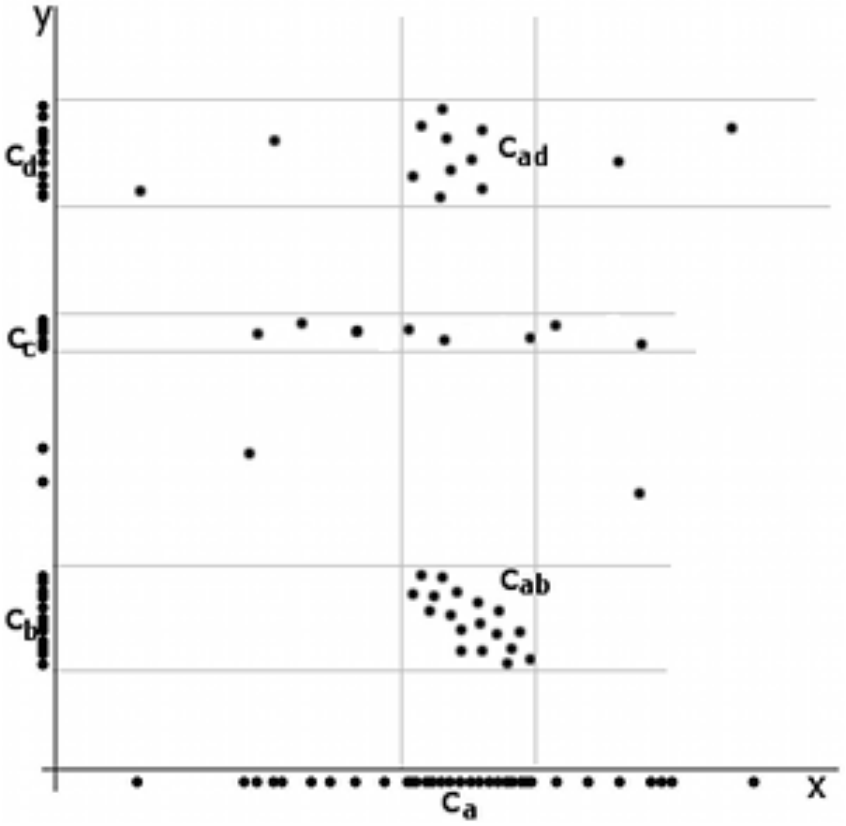


Fig. 7. Subspace Clustering [Slowmo 2010]

correlated. 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 [Madeira and Oliveira 2004].

Examples are as follows. CTWC (Coupled Two-Way Clustering) [Getz et al. 2000] 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) [Tang et al. 2001] clusters the rows and then clusters the columns, based on each row cluster. It keeps the cluster pairs that are most dissimilar. Block Clustering [Hartigan 1975] 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. $\delta$-biclusters [Cheng and Church 2000] 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) [Tanay et al. 2004] uses probabilistic modeling and graph theoretic techniques to find subsets of rows whose values are very different in a subset of columns. Plaid Models [Lazzeroni and Owen 2000] allows biclusters to overlap, *i.e.* a gene or a sample/condition can belong to more than one cluster. Information-theoretic co-clustering [Dhillon et al. 2003] intertwines the row and column clusterings to increase mutual information.

*3.8.5. 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) [Aggarwal and Yu 2000] 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) [Böhm et al. 2004] 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) [Achtert et al. 2006] 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) [Achtert et al. 2008] is based on the Hough transform [Hough 1962], 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.

### 3.9. Other Clustering Techniques

*3.9.1. Neural Network-Based Clustering:.* The neural network approach has been studied intensively by mathematicians, statisticians, physicists, engineers, and computer scientists [Kumar and Joshi 2007]. 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 [Du 2010]. Competitive learning can be implemented using a two-layer neural network: the input layer and the output layer [Du 2010].

Examples are as follows. A SOM (Self-Organizing Map) [Kohonen 1982] 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) [Burton et al. 1983; Gersho and Ramamurthi 1982] 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) [Gersho and Ramamurthi 1982]. Neural gas [Martinetz and Schulten 1991] 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. ART model is covered in Section 3.7.4.

*3.9.2. 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 selec-

tion 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 [Fogel 2006] provide an alternate and effective way to find the global or approximately global optimum [Xu and Wunsch 2010]. PSO (Particle Swarm Optimization) simulates social behavior in nature, such as bird flocking or fish schooling [Kennedy et al. 2001]. ACO (Ant Colony Optimization) algorithms model the behaviors of ants in nature [Dorigo and Stützle 2003]. GAs (Genetic Algorithms) [Holland 1975] mimic natural selection and use evolutional mechanisms such as crossover, mutation and selection to generate solutions.

Examples are as follows. PPO (Particle-Pair Optimizer) [Du et al. 2008] 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 [Sheng et al. 2004] modifies Deterministic Crowding [Mahfoud 1995], one of the niching genetic algorithms, and incorporates one step of k-means into its regeneration steps [Sheng et al. 2004]. EvoCluster algorithm [Ma et al. 2006] 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 [Ges et al. 2005] 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.

*3.9.3. Kernel Clustering:.* Kernel-based learning such as Support Vector Machines (SVMs) [Cortes and Vapnik 1995; Schölkopf et al. 1996; Platt 1999] has had successful applications in pattern recognition and machine learning and is becoming increasingly important [Platt 1999]. Kernel methods [Cristianini and Shawe-Taylor 2000] perform a nonliner 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) [Wang et al. 2007; Zafeiriou and Laskaris 2008] uses SVM training to find the cluster boundaries and an adjacency matrix to assign a cluster label to each data object [Xu and Wunsch 2010]. Variations of SVC include Iterative One-Class SVC [Camastra and Verri 2005], and rough Set SVC [Pawlak 1991]. Kernel k-means [Girolami 2002] 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 [Schölkopf et al. 1998]. Kernel deterministic annealing clustering [Yang and Wu 2006] uses an adaptively selected Gaussian parameter and a Gaussian kernel to determine the nonlinear mapping. Kernel fuzzy clustering [Liu and Xu 2008; Zhang and Chen 2004; Zhou and Gan 2004] applies kernel techniques to fuzzy clustering algorithms by replacing the original Euclidean distance with a kernel-induced distance. Kernel Self-Organizing Maps [Andras 2002; Boulet et al. 2008; Lau et al. 2006] 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.

*3.9.4. 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 [Gusfield 1997; Sun and Giles 2001], 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 [Xu and Wunsch 2010]. 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 [Xu and Wunsch 2010]. Examples are as follows. The Needleman-Wunsch algorithm [Durbin et al. 1998; Needleman and Wunsch 1970] uses basic dynamic programming and is a global optimal alignment algorithm. The Smith-Waterman algorithm [Durbin et al. 1998; Smith and Waterman 1980] 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) [Pearson and Lipman 1988] 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) [Altschul et al. 1990] searches for short alignment matches between two sequences using a heuristic approach, which approximates the Smith-Waterman algorithm. GeneRage [Enright and Ouzounis 2000] automatically clusters sequence datasets by using Smith-Waterman dynamic programming alignment and single-linkage clustering. SEQOPTICS (SEQuence clustering with OPTICS) [Chen et al. 2006] implements Smith-Waterman algorithms as the distance measurement and uses OPTICS [Ankerst et al. 1999] 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 [Xu and Wunsch 2010]. Examples are as follows. Scalable sequential data clustering [Guralnik and Karypis 2001] uses a k-means based clustering algorithm which has near-linear time complexity to improve the scalability problem. Pattern-oriented hierarchical clustering [Morzy et al. 1999] 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 [Vlachos et al. 2003] 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 [Xu and Wunsch 2010]. Examples are as follows. Autoregressive moving average (ARMA) models [Bagnall and Janacek 2004; Xiong and Yeung 2004] 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 [Ramoni et al. 2002; Smyth 1999] 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 [Bagnall et al. 2003; Gaffney and Smyth 1999] 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) [Oates et al. 2001; Smyth 1996] 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.

*3.9.5. Ensemble Clustering:.* Clustering ensembles have emerged to improve robustness, stability and accuracy of clustering results [Ghaemi et al. 2009]. A cluster ensemble combines the results of multiple clustering algorithms to obtain a consensus result [Pirim et al. 2011]. 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 [Jain 2007].

There are two main steps in a clustering ensemble: generation and consensus. In the generation step, several approaches are used [Vega-Pons and Ruiz-Shulcloper 2011]: 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 [Dudoit and Fridlyand 2003] 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 [Dudoit and Fridlyand 2003] 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 [Azimi et al. 2007] 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 [Luo et al. 2006] uses a GA to find a combined clustering by minimizing an information-theoretical criterion function. The generalized MI clustering ensemble [Topchy et al. 2003] 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 [Kellam et al. 2001] 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 [Fred 2001] transforms data partitions into a co-association matrix with coherent association mappings. It then extracts underlying clusters from this matrix. Evidence accumulation-based clustering [Fred and Jain 2002] 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' [Vega-Pons and Ruiz-Shulcloper 2011]. It obtains the consensus clustering result by solving a maximum likelihood estimation problem. Mixture model clustering ensemble [Topchy et al. 2004] uses a probabilistic model of consensus based

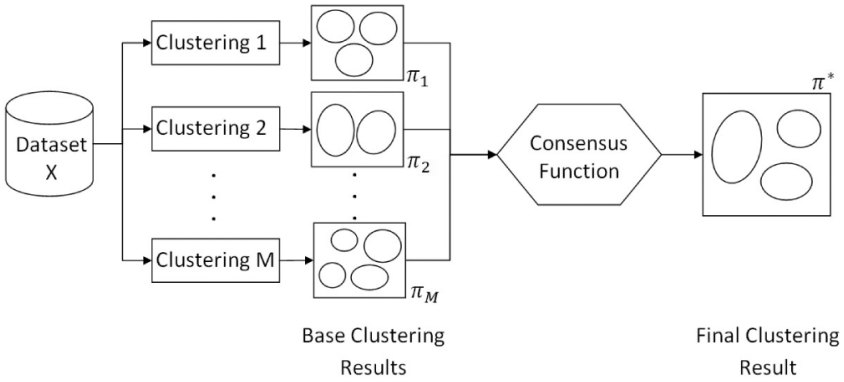


Fig. 8. Ensemble Clustering [Iam-On et al. 2012]

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 [Vega-Pons and Ruiz-Shulcloper 2011]. Examples are as follows. METIS [Karypis and Kumar 1998] 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) [Ng et al. 2001] 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) [Strehl and Ghosh 2002] first creates a graph based on a co-association matrix, and then performs METIS clustering on the graph. HGPA (Hypergraph Partitioning Algorithm) [Strehl and Ghosh 2002] uses a hyperedge in a graph to represent each cluster. It then uses minimal cut algorithms such as HMETIS [Karypis et al. 1999] to find good hypergraph partitions. MCLA (Meta Clustering Algorithm) [Strehl and Ghosh 2002] determines soft cluster membership values for each data object by using hyperedge collapsing operations. HBGF (Hybrid Bipartite Graph Formulation) [Fern and Brodley 2003] 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.

Other approaches are as follows. The cumulative voting consensus method [Ayad and Kamel 2008] solves the cluster label alignment problem by using cumulative voting, where a probabilistic mapping between labels is computed. Bipartite Merger and Metis merger [Hore et al. 2009] are approaches for merging an ensemble of clustering solutions using sets of cluster centers. They are highly scalable and provide competitive results. Weighted consensus clustering [Li and Ding 2008] weights each input clustering. It determines weights in a way so that the clusters are better separated. Bayesian Cluster Ensembles [Wang et al. 2011] takes a Bayesian approach to combine clusterings. It uses a variational approximation based algorithm for learning. This way, it is able to avoid the cluster label correspondence problems.

Figure 8 shows an example of ensemble clustering.

*3.9.6. 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 [Ferligoj and Batagelj 1992].

Examples are as follows. FCPSO (Fuzzy Clustering-based Particle Swarm Optimization) [Agrawal et al. 2008] uses an external repository to save nondominated 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 [Handl and Knowles 2004] and MOCK (MultiObjective Clustering with automatic k-determination) [Handl and Knowles 2007] 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) [Corne et al. 2001] to optimize two complementary clustering objectives. Multi-objective real coded genetic fuzzy clustering [Mukhopadhyay and Maulik 2011] 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) [Romero-Zliz et al. 2008] combines evolutionary algorithms with multi-objective optimization techniques and relies on the NSGA-II multi-objective genetic algorithm [Deb et al. 2002]. It can discover less obvious but informative data associations.

*3.9.7. Semi-supervised Clustering:.* Semi-supervised clustering provides limited supervision to unsupervised clustering. There are many cases when some knowledge about the data is available such as the constraints between data objects or cluster labels for some data objects. Such knowledge can be used to guide the clustering process. There are several approaches for semi-supervised clustering: similarity-adapting methods, search-based methods, and other methods.

Similarity-adapting methods use a similarity measure which is adapted to make the available constraints more easily satisfied [Grira et al. 2005]. Examples are as follows. Distance metric learning based clustering [Xing et al. 2003] learns a distance metric based on examples of similar pairs of data objects in the input space using convex optimization. Space-level constraints based clustering [Klein et al. 2002] exploits space-level implications based on instance-level constraints. It uses an all-pairs-shortest-paths algorithm to adjust the distance metric.

Search-based methods modify the clustering algorithm itself to use the available constraints or labels to guide the search for an appropriate clustering [Grira et al. 2005]. Examples are as follows. Seeded-K Means and Constrained-K Means [Basu et al. 2002] generate initial seed clusters based on labeled data. The latter also generates constraints from labeled data and guides the clustering process using those constraints. Semi-Supervised Clustering Using Genetic Algorithms [Demiriz et al. 1999] modifies k-means clustering to minimize within-cluster variance and a measure of cluster impurity. Clustering with Instance-level Constraints [Wagstaff and Cardie 2000] incorporates hard constraints using a modified version of Cobweb (covered in Section 3.7.4) which partitions the data.

Other methods include the probabilistic semi-supervised clustering with constraints method [Basu et al. 2006], which derives an objective function from the joint probability defined over the Hidden Markov Random Field model and performs semi-supervised clustering by minimizing this object function.

## 4. APPLICATIONS OF CLUSTERING IN CANCER SUBTYPING

The recently-developed DNA microarray technologies [Eisen and Brown 1999; Lipshutz et al. 1999], which can measure the expression levels of tens of thousands of genes simultaneously, offer cancer researchers a novel method 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 [McLachlan et al. 2005; Xu and Wunsch 2005]. Investigations on leukemia [Golub et al. 1999], lymphoma [Alizadeh et al. 2000], colon cancer [Alon et al. 1999], cutaneous melanoma [Bittner et al. 2000], bladder cancer [Dyrskjøt et al. 2002], breast cancer [Perou et al. 2000], lung cancer [Garber et al. 2001], and others show very promising results. Supervised computational methods, such as multi-layer perceptrons [Khan et al. 2001], naive Bayes [Li et al. 2004a], support vector machines [Li et al. 2004b; Statnikov et al. 2005], semi-supervised Ellipsoid ARTMAP [Xu et al. 2007], and k-Top Scoring Paris [Tan et al. 2005], 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 [Xu and Wunsch 2005] 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.

Microarray data sets are generally generated from two types of platforms, single-channel microarrays (*e.g.* Affymetrix) or double-channel microarrays (*e.g.* cDNA) [Quackenbush 2001; Monti et al. 2003; Slonim 2002]. Measurements of Affymetrix arrays are based on the number of RNA copies found in the cell sample, and measurements of cDNA microarrays are based on the ratios of the number of copies in relation to a control cell sample [De Souto et al. 2008]. Microarray data include mRNA gene expression data and miRNA gene expression data.

Subtyping by microarrays is challenging due to the following reasons: a) There are a large number of genes, but a relatively small number of 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. The number of genes are usually in the thousands range, while the sample sizes are usually in the tens range. 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 [Berkhin 2002; Beyer et al. 1999]. This effect becomes much worse when dimensions are greater than 15, and proximity based clustering becomes unreliable [Berkhin 2002]. b) It is important to identify the most relevant genes. 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 [Chowers et al. 2003]. And they can be classified into three major categories: biological - cells from different populations, tissues, conditions, etc. experimental - defects of the spotting equipment, different hybridiza-

tion 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 [Zhou 2003].

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 [O'Day and Lal 2010]. miRNAs are small non-protein coding RNAs found in animals and plants. The first miRNAs were discovered and characterized in the early 1990s [Lee et al. 1993]. 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 [Calin et al. 2002]. Later on many miRNAs have been found to be related to more types of cancer [Bottoni et al. 2005; Chan et al. 2005; Ciafre et al. 2005; Iorio et al. 2005; Johnson et al. 2005; Karube et al. 2005; Metzler et al. 2004; Michael et al. 2003; Yanaihara et al. 2006]. 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.

### 4.1. Clinical Applications

*4.1.1. mRNA-based Applications.* Golub used mRNA profiling of the expression of 6,817 genes in 72 leukemia samples as a test case for subtyping [Golub et al. 1999]. 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 [Alizadeh et al. 2000]. 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 [Armstrong et al. 2002]. 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 [Bittner et al. 2000]. 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 [Maniotis et al. 1999]. 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 [Perou et al. 2000]. Three subtypes were discovered in this clinically highly heterogeneous tumor: the previously known Erb-B2, and two others previously unknown, namely ER+(estrogen receptorpositive)/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 [Lapointe et al. 2004]. 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 [Liang et al. 2005]. 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 [Laiho et al. 2006]. 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 [Monti et al. 2003; Wilkerson et al. 2010; Wilkerson and Hayes 2010]. 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 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 [Lei et al. 2013]. 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.

*4.1.2. microRNA-based Applications.* Lu performed computational analyses on 217 miRNAs from 334 mammalian samples, including multiple human cancers [Lu et al. 2005]. 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 [Blenkiron et al. 2007]. 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+ [Mattie et al. 2006]. 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 [De Laurentiis et al. 2005]. 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 [Porkka et al. 2007]. 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 [Oberg et al. 2011]. 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 [Yang et al. 2013]. 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.

### 4.2. Computational Experiments

*4.2.1. mRNA-based Experiments.* Xing and Karp presented CLIFF (CLustering via Iterative Feature Filtering) [Xing and Karp 2001] and tested it with mRNA profiles of 72 leukemia samples and 7130 genes [Golub et al. 1999]. 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 clus-

ters 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 [Tang and Zhang 2002], and tested it on mRNA profiles of over 7129 genes in 72 leukemia patient samples [Golub et al. 1999]. 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 [Getz et al. 2000] (covered in Section 3.8.4) and applied it to a leukemia mRNA data set [Golub et al. 1999] and a colon cancer mRNA data set [Alon et al. 1999]. 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 [Cheng and Church 2000] and applied it to diffuse large B-cell lymphoma mRNA profiles containing 4026 genes and 96 conditions [Alizadeh et al. 2000]. 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) [Iam-on et al. 2010] and applied it to several mRNA profile data sets including: leukemia1 [Golub et al. 1999] (1,877 genes and 72 samples), leukemia2 [Golub et al. 1999] (1,877 genes and 72 samples), leukemia3 [Armstrong et al. 2002] (2,194 genes and 72 samples), brain tumor [Nutt et al. 2003] (1,377 genes and 50 samples), central nervous system (CNS) [Pomeroy et al. 2002] (1,379 genes and 42 samples), and hepatocellular carcinoma (HCC) [Chen et al. 2002] (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 [Kim et al. 2009], consensus clustering with hierarchical clustering [Monti et al. 2003], graph-based consensus clustering [Yu et al. 2007], Cluster based Similarity Partitioning Algorithm [Strehl and Ghosh 2002], Hyper-Graph Partitioning Algorithm [Strehl and Ghosh 2002], Meta-Clustering Algorithm [Strehl and Ghosh 2002], Hybrid Bipartite Graph Formulation [Fern and Brodley 2004], 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.

*4.2.2. microRNA-based Experiments.* Lock and Dunson proposed Bayesian Consensus Clustering (BCC) [Lock and Dunson 2013] and tested it with miRNA profiles of 348

breast cancer samples and 423 miRNAs [Koboldt et al. 2012]. 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) [Li et al. 2014] and tested it with the miRNA profiles of 71 breast cancer patients and 13 miRNAs [Blenkiron et al. 2007]. 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 [Blenkiron et al. 2007] were identified successfully by CTWC. The clinical significance of the identified subtypes were verified using Kaplan-Meier survival analysis [Kopycka-Kedzierawski and Billings 2004].

## 5. CHALLENGES

Advances in microarray 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 [Van't Veer and Bernards 2008]. However, there are still many challenges.

### 5.1. 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 [Perou et al. 1999]. 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.

### 5.2. 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.

**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 [Ramaswamy and Golub 2002]. 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 [Ramaswamy and Golub 2002]. 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.

## 6. CONCLUSIONS AND FUTURE PERSPECTIVES

### 6.1. Conclusions

In this paper 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 [Kuncheva and Hadjitodorov 2004]. The performance of a given clustering algorithm and a validation matrix is dependent on data characteristics and the application [Jiang et al. 2004].

### 6.2. Future Perspectives

**Different granularities** Users often desire different cluster granularity for different 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.

**Interactive semi-supervised feature** Prior to clustering, much information about the data set is unknown, but some knowledge is often available. For example, some experimental conditions are known to be correlated. If a clustering algorithm can provide the options to interactively integrate such partially available knowledge into the clustering process, it may produce more biologically meaningful results [Pirim et al. 2012].

**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 hyperspheres 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.

## REFERENCES


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

Elke Achtert, Christian Böhm, Peer Kröger, and Arthur Zimek. 2006. Mining hierarchies of correlation clusters. In *Scientific and Statistical Database Management, 2006. 18th International Conference on*. IEEE, 119–128.

Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. 1999. Fast algorithms for projected clustering. In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, USA, 61–72. DOI:http://dx.doi.org/10.1145/304182.304188

Charu C. Aggarwal and Philip S. Yu. 2000. Finding generalized projected clusters in high dimensional spaces. *SIGMOD Rec.* 29, 2 (May 2000), 70–81. DOI:http://dx.doi.org/10.1145/335191.335383

Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. 1998. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *SIGMOD Rec.* 27, 2 (June 1998), 94–105. DOI:http://dx.doi.org/10.1145/276305.276314

Shubham Agrawal, B. K. Panigrahi, and Manoj Kumar Tiwari. 2008. Multiobjective Particle Swarm Algorithm With Fuzzy Clustering for Electrical Power Dispatch. *IEEE Trans. Evolutionary Computation* 12, 5 (2008), 529–541. http://dblp.uni-trier.de/db/journals/tec/tec12.html#AgrawalPT08

Khaled S. Al-Sultan and Shokri Z. Selim. 1993. A global algorithm for the fuzzy clustering problem. *Pattern Recognition* 26, 9 (1993), 1357–1361. http://dblp.uni-trier.de/db/journals/pr/pr26.html#Al-SultanS93

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2007. *Molecular Biology of the Cell* (5th ed.). Garland Science, Taylor & Francis Group, New York.

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. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 6769 (3 Feb. 2000), 503–511. DOI:http://dx.doi.org/10.1038/35000501

Derek C. Allen. 2013. *e-Study Guide for Histopathology Specimens: Clinical, Pathological and Laboratory Aspects*. Cram101 Textbook Reviews. https://books.google.com/books?id=v5nv8E-gUh4C

Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. 1999. 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 (1999), 6745–6750.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215, 3 (1990), 403 – 410. DOI:http://dx.doi.org/10.1016/S0022-2836(05)80360-2

G. C. Anagnostopoulos and M. Georgiopoulos. 2001. Ellipsoid ART and ARTMAP for incremental clustering and classification. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, Vol. 2. 1221–1226 vol.2. DOI:http://dx.doi.org/10.1109/IJCNN.2001.939535

Peter Andras. 2002. Kernel-Kohonen Networks. *Int. J. Neural Syst.* 12, 2 (2002), 117–135. http://dblp.uni-trier.de/db/journals/ijns/ijns12.html#Andras02

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: Ordering points to identify the clustering structure. *SIGMOD Rec.* 28, 2 (June 1999), 49–60. DOI:http://dx.doi.org/10.1145/304181.304187

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. 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics* 30, 1 (2002), 41–47.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1027–1035. http://dl.acm.org/citation.cfm?id=1283383.1283494

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

Hanan G Ayad and Mohamed S Kamel. 2008. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1 (2008), 160–173.

J. Azimi, M. Mohammadi, A. Movaghar, and M. Analoui. 2007. Clustering Ensembles Using Genetic Algorithm. In *Computer Architecture for Machine Perception and Sensing, 2006. CAMP 2006. International Workshop on*. 119 –123. DOI:http://dx.doi.org/10.1109/CAMP.2007.4350366

G. Phanendra Babu and M. Narasimha Murty. 1993. A near-optimal initial seed value selection in K-means means algorithm using a genetic algorithm. *Pattern Recognition Letters* 14, 10 (1993), 763–769. http://dblp.uni-trier.de/db/journals/prl/prl14.html#BabuM93

A. Bagnall, G. Janacek, B. Iglesia, and M. Zhang. 2003. Clustering time series from mixture polynomial models with discretized data. *Proc. 2nd Australasian Data Mining Workshop* (2003), 105120.

Anthony J. Bagnall and Gareth J. Janacek. 2004. Clustering time series from ARMA models with clipped data.. In *KDD* (2006-02-13), Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel (Eds.). ACM, 49–58. http://dblp.uni-trier.de/db/conf/kdd/kdd2004.html#BagnallJ04

G. H. Ball and D. J. Hall. 1965. ISODATA, an iterative method of multivariate analysis and pattern classification. *IFIPS Congress* (1965).

A. Baraldi and E. Alpaydin. 2002. Constructive feedforward ART clustering networks – Part I and II. *IEEE Trans. Neural Netw.* 13, 3 (May 2002).

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

Daniel Barbara, Julia Couto, and Yi Li. 2002. COOLCAT: An entropy-based algorithm for categorical clustering. In *In Proceedings of the Eleventh International Conference on Information and Knowledge Management*. ACM Press, 582–589.

John M. S. Bartlett and David Stirling. 2003. A short history of the polymerase chain reaction. In *PCR Protocols*. Springer, 3–6.

S. Basu, A. Banerjee, and R. Mooney. 2002. Semi-supervised clustering by seeding. In *Proceedings of the International Conference on Machine Learning*. 27–34.

Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J Mooney. 2006. Probabilistic semi-supervised clustering with constraints. *Semi-Supervised Learning* (2006), 71–98.

Stephen D. Bay and Michael J. Pazzani. 1999. 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)*. ACM, New York, NY, USA, 302–306. DOI:http://dx.doi.org/10.1145/312129.312263

Ron Bekkerman, Ran El-Yaniv, and Andrew McCallum. 2005. Multi-way distributional clustering via pairwise interactions. In *Proceedings of the 22nd International Conference on Machine learning*. ACM, 41–48.

Mikhail Belkin and Partha Niyogi. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 15, 6 (2003), 1373–1396.

Pavel Berkhin. 2002. *Survey Of Clustering Data Mining Techniques*. Technical Report. Accrue Software, San Jose, CA. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.3739

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is nearest neighbor meaningful? In *Database Theory ICDT99*. Springer, 217–235.

J. Bezdek, C. Coray, R. Gunderson, and J. Watson. 1981. Detection and Characterization of Cluster Substructure I. Linear Structure: Fuzzy c-Lines. *SIAM J. Appl. Math.* 40, 2 (1981), 339–357. DOI:http://dx.doi.org/10.1137/0140029

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. 2001. 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 (1 Nov. 2001), 13790–13795. DOI:http://dx.doi.org/10.1073/pnas.191502998

D. A. Binder. 1978. Bayesian cluster analysis. *Biometrika* 65, 1 (1978), 31–38. http://biomet.oxfordjournals.org/cgi/reprint/65/1/31.pdf

C. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford Univ. Press.

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. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406, 6795 (3 Aug. 2000), 536–40.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

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, and others. 2007. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol* 8, 10 (2007), R214–R229.

Hans H. Bock. 1996. Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis* 23, 1 (15 Nov. 1996), 5–28. http://www.sciencedirect.com/science/article/B6V8V-3SWT44Y-2/1/04b59331d005358183bffbdb5b9b3c6e

Christian Böhm, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. 2004. Density Connected Clustering with Local Subspace Preferences.. In *ICDM* (2004-12-01). IEEE Computer Society, 27–34. http://dblp.uni-trier.de/db/conf/icdm/icdm2004.html#BohmKKK04

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

Romain Boulet, Bertrand Jouve, Fabrice Rossi, and Nathalie Villa. 2008. Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing* 71, 7-9 (2008), 1257–1273. http://dblp.uni-trier.de/db/journals/ijon/ijon71.html#BouletJRV08

Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. 2010. Random Projections for k-means Clustering.. In *Advances in Neural Information Processing Systems*. 298–306.

Paul S. Bradley, Usama M. Fayyad, and Cory Reina. 1998. Scaling Clustering Algorithms to Large Databases. In *Knowledge Discovery and Data Mining*. 9–15. citeseer.ist.psu.edu/bradley98scaling.html

Nathan Brannon, John Seiffertt, Timothy Draelos, and Donald C. Wunsch II. 2009. Coordinated machine learning and decision support for situation awareness. *Neural Networks* 22, 3 (2009), 316–325. http://dblp.uni-trier.de/db/journals/nn/nn22.html#BrannonSDW09

Markus Bredel, Claudia Bredel, Dejan Juric, Griffith R. Harsh, Hannes Vogel, Lawrence D. Recht, and Branimir I. Sikic. 2005. Functional network analysis reveals extended gliomagenesis pathway maps and three novel Myc-interacting genes in human gliomas. *Cancer Research* 65, 19 (2005), 8679–8689.

Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* 101, 12 (23 March 2004), 4164–4169. DOI:http://dx.doi.org/10.1073/pnas.0308531101

D. Burton, J. Shore, and J. Buck. 1983. A generalization of isolated word recognition using vector quantization. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, Vol. 8. 1021–1024. DOI:http://dx.doi.org/10.1109/ICASSP.1983.1171915

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. 2002. 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 (2002), 15524–15529.

Francesco Camastra and Alessandro Verri. 2005. A Novel Kernel Method for Clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27, 5 (2005), 801–804. http://dblp.uni-trier.de/db/journals/pami/pami27.html#CamastraV05

Gail A. Carpenter and Stephen Grossberg. 1987a. ART 2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns. *Applied Optics* 26, 23 (1987), 4919–4930.

Gail A. Carpenter and Stephen Grossberg. 1987b. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing* 37, 1 (1987), 54–115.

G. A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds, and D.B. Rosen. 1992. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Transactions on Neural Networks* 3, 5 (September 1992), 698–713.

Gail A. Carpenter, Stephen Grossberg, and John H. Reynolds. 1991. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks* 4, 5 (1991), 565–588.

Thomas P Caudell, Scott DG Smith, G Craig Johnson, and Donald C Wunsch II. 1991. Application of neural networks to group technology. In *Applications of Artificial Neural Networks II*. International Society for Optics and Photonics, 612–621.

Soumen Chakrabarti. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.

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

Jae-Woo Chang and Du-Seok Jin. 2002. A new cell-based clustering method for large, high-dimensional data in data mining applications. In *Proceedings of the 2002 ACM symposium on Applied computing*. ACM, 503–507.

Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. 1997. Incremental Clustering and Dynamic Information Retrieval. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing (STOC '97)*. ACM, New York, NY, USA, 626–635. DOI:http://dx.doi.org/10.1145/258533.258657

Peter Cheeseman and John Stutz. 1996. Advances in knowledge discovery and data mining. In *Bayesian Classification (AutoClass): Theory and Results*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA, 153–180. http://dl.acm.org/citation.cfm?id=257938.257954

Farid F Chehab. 1993. Molecular diagnostics: Past, present, and future. *Human Mutation* 2, 5 (1993), 331–337.

Jiun-Rung Chen. 2010. *Efficient Biclustering Methods for Microarray Databases*. Ph.D. Dissertation. National Sun Yat-sen University.

Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y Chang. 2011. Parallel spectral clustering in distributed systems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33, 3 (2011), 568–586.

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. 2002. Gene Expression Patterns in Human Liver Cancers. *Molecular Biology of the Cell* 13, 6 (2002), 1929–1939.

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

Chun Hung Cheng, Ada Wai-Chee Fu, and Yi Zhang. 1999. Entropy-based Subspace Clustering for Mining Numerical Data.. In *KDD*, Usama M. Fayyad, Surajit Chaudhuri, and David Madigan (Eds.). ACM, 84–93. http://dblp.uni-trier.de/db/conf/kdd/kdd99.html#ChengFZ99

Yizong Cheng and George M. Church. 2000. Biclustering of Expression Data. In *Proc. of the 8th Intelligent Systems for Molecular Biology*. AAAI Press, 93–103. citeseer.nj.nec.com/cheng00biclustering.html

Chire. 2011a. Density-Based Clustering with DBSCAN. (2011). URL http://commons.wikimedia.org/wiki/File:DBSCAN-density-data.svg. Accessed Oct. 2015.

Chire. 2011b. Expectation-Maximization (EM) Clustering Examples. (2011). URL http://commons.wikimedia.org/wiki/File:EM-Gaussian-data.svg. Accessed Oct. 2015.

Chire. 2011c. k-Means Clustering Examples. (2011). URL http://commons.wikimedia.org/wiki/File:KMeans-Gaussian-data.svg. Accessed Oct. 2015.

D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder. 2006. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of Molecular Diagnostics* 8, 1 (Feb. 2006), 31–39. http://view.ncbi.nlm.nih.gov/pubmed/16436632

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. 2003. Gene expression variation in the adult human retina. *Human Molecular Genetics* 12, 22 (2003), 2881–2893.

S. A. Ciafre, S. Galardi, A. Mangiola, M. Ferracin, C.-G. Liu, G. Sabatino, M. Negrini, G. Maira, C. M. Croce, and M. G. Farace. 2005. Extensive modulation of a set of microRNAs in primary glioblastoma. *Biochemical and Biophysical Research Communications* 334, 4 (2005), 1351–1358.

Albert H. Coons, Hugh J. Creech, R. Norman Jones, and Ernst Berliner. 1942. The demonstration of pneumococcal antigen in tissues by the use of fluorescent antibody. *The Journal of Immunology* 45, 3 (1942), 159–70.

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

David W Corne, Nick R Jerram, Joshua D Knowles, Martin J Oates, and others. 2001. PESA-II: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*.

Carlos D Correa and Peter Lindstrom. 2012. Locally-scaled spectral clustering using empty region graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1330–1338.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine learning* 20, 3 (1 Sept. 1995), 273–297. DOI:http://dx.doi.org/10.1023/a:1022627411411

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (1st ed.). Cambridge University Press. http://www.cs.orst.edu/~bulatov/papers/CambridgeUniversityPress-Support.Vector.Machines.and.Oth.chm

R. D'Andrade. 1978. U-statistic hierarchical clustering. *Psychometrika* 4 (1978).

M. Dash, H. Liu, and Xiaowei Xu. 2001. '1+1 $>$ 2': Merging distance and density based clustering. In *Database Systems for Advanced Applications, 2001. Proceedings. Seventh International Conference on*. IEEE Computer Society, 32 –39. DOI:http://dx.doi.org/10.1109/DASFAA.2001.916361

N. E. Day. 1969. Estimating the components of a mixture of normal distributions. *Biometrika* 56, 3 (1969), 463–474. DOI:http://dx.doi.org/10.1093/biomet/56.3.463

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

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

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

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

Ayhan Demiriz, Kristin Bennett, and Mark J. Embrechts. 1999. Semi-Supervised Clustering Using Genetic Algorithms. In *In Artificial Neural Networks in Engineering (ANNIE-99)*. ASME Press, 809–814.

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

P. D'haeseleer. 2005. How does gene expression clustering work? *Nature Biotechnology* 23, 12 (2005), 1499–1501.

I. Dhillon and D. Modha. 1999. A data clustering algorithm on distributed memory multiprocessors. In *5th ACM SIGKDD, Large-scale Parallel KDD Systems Workshop*. 245–260.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. 2004. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 551–556.

Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. 2003. Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 89–98.

L. J. DiDio. 1994. Marcello Malpighi: The father of microscopic anatomy. *Italian Journal of Anatomy and Embryology* 100 (1994), 3–9.

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

K.-L. Du. 2010. Clustering: A neural network approach. *Neural Networks* 23, 1 (2010), 89–107. http://dblp.uni-trier.de/db/journals/nn/nn23.html#Du10

Zhihua Du, Yiwei Wang, and Zhen Ji. 2008. PK-means: A new algorithm for gene clustering. *Computational Biology and Chemistry* 32, 4 (2008), 243–247. http://dblp.uni-trier.de/db/journals/candc/candc32.html#DuWJ08

R. O. Duda, P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*. Wiley. http://books.google.com/books?id=YoxQAAAAMAAJ

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

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

Jeyapradha Duraiyan, Rajeshwar Govindarajan, Karunakaran Kaliyappan, and Murugesan Palanisamy. 2012. Applications of immunohistochemistry. *Journal of Pharmacy & Bioallied Sciences* 4, Suppl 2 (2012), S307.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. http://www.amazon.com/Biological-Sequence-Analysis-Probabilistic-Proteins/dp/0521629713

L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Orntoft. 2002. Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics* 33, 1 (2002), 90–96.

A. W. F. Edwards and L. L. Cavalli-Sforza. 1965. A method for cluster analysis. *Biometrics* 21 (1965), 362–375.

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

C. Eldershaw and M. Hegland. 1997. Cluster analysis using triangulation. *Computational Techniques and Applications* (1997), 201–208.

Charles Elkan. 2003. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, Vol. 3. 147–153.

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

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

Martin Ester, Hans P. Kriegel, Jorg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second International Conference on Knowledge Discovery and Data Mining*, Evangelos Simoudis, Jiawei Han, and Usama Fayyad (Eds.). AAAI Press, Portland, Oregon, 226–231. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2930

Tim Evans. 2014. Zachary Karate Club network clustered using clique graph methods. (2014). URL http://netplexity.org/?p=1261. Accessed Oct. 2015.

B. Everitt, S. Landau, and M. Leese. 2001. *Cluster Analysis*. London: Arnold.

B. S. Everitt. 1980. Cluster Analysis. In *Cluster Analysis, Second Edition*. Heineman Educational Books Ltd.

Christos Faloutsos and King-Ip Lin. 1995. FastMap: A Fast Algorithm for Indexing, Data-mining and Visualization of Traditional and Multimedia Datasets. *SIGMOD Rec.* 24, 2 (May 1995), 163–174. DOI:http://dx.doi.org/10.1145/568271.223812

A. Ferligoj and V. Batagelj. 1992. Direct multicriterion clustering algorithms. *Journal of Classification* 9 (1992), 43–61.

Xiaoli Zhang Fern and Carla E. Brodley. 2003. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach.. In *International Conference on Machine Learning*, Tom Fawcett and Nina Mishra (Eds.). AAAI Press, 186–193. http://dblp.uni-trier.de/db/conf/icml/icml2003.html#FernB03a

Xiaoli Zhang Fern and Carla E Brodley. 2004. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the Twenty-First International Conference on Machine Learning*. ACM, 36.

Douglas H Fisher. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning* 2, 2 (1987), 139–172.

David B Fogel. 2006. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. John Wiley & Sons.

E. W. Forgy. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* 21 (1965), 768–769.

Ildiko E. Frank and Roberto Todeschini. 1994. *The Data Analysis Handbook*. Elsevier Science Inc.

Ana L. N. Fred. 2001. Finding Consistent Clusters in Data Partitions.. In *Multiple Classifier Systems (Lecture Notes in Computer Science)*, Josef Kittler and Fabio Roli (Eds.), Vol. 2096. Springer, 309–318. http://dblp.uni-trier.de/db/conf/mcs/mcs2001.html#Fred01

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

Scott Gaffney and Padhraic Smyth. 1999. Trajectory Clustering with Mixtures of Regression Models. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 63–72.

Guojun Gan. 2011. *Data Clustering in C++: An Object-Oriented Approach*. Taylor & Francis. https://books.google.com/books?id=5ScnAlm51tgC

Guojun Gan, Chaoqun Ma, and Jianhong Wu. 2007. *Data Clustering - Theory, Algorithms, and Applications*. SIAM. I–XXII, 1–466 pages.

Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. 1999a. CACTUS - Clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99)*. ACM, New York, NY, USA, 73–83. DOI:http://dx.doi.org/10.1145/312129.312201

Venkatesh Ganti, Raghu Ramakrishnan, Johannes Gehrke, Allison Powell, and James French. 1999b. Clustering large datasets in arbitrary metric spaces. In *Data Engineering, 1999. Proceedings., 15th International Conference on*. IEEE, 502–511.

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. 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences* 98, 24 (20 November 2001), 1378413789. DOI:http://dx.doi.org/10.1073/pnas.241500798

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

Vito Di Ges, Raffaele Giancarlo, Giosu Lo Bosco, Alessandra Raimondi, and Davide Scaturro. 2005. GenClust: A genetic algorithm for clustering gene expression data. *BMC Bioinformatics* 6, 1 (2005), 289–289. http://dblp.uni-trier.de/db/journals/bmcbi/bmcbi6.html#GesuGBRS05

Gad Getz, Erel Levine, and Eytan Domany. 2000. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences* 97, 22 (2000), 12079–12084. DOI:http://dx.doi.org/10.1073/pnas.210134797

Reza Ghaemi, Md. Nasir Sulaiman, Hamidah Ibrahim, and Norwati Mustapha. 2009. A Survey: Clustering Ensembles Techniques. *World Academy of Science, Engineering and Technology* 50 (2009), 636–645.

David Gibson, Jon Kleinberg, and Prabhakar Raghavan. 2000. Clustering categorical data: An approach based on dynamical systems. *The International Journal on Very Large Data Bases* 8, 3-4 (Feb. 2000), 222–236. DOI:http://dx.doi.org/10.1007/s007780050005

M. Girolami. 2002. Mercer kernel-based clustering in feature space. *Neural Networks, IEEE Transactions on* 13, 3 (07 Aug. 2002), 780–784. DOI:http://dx.doi.org/10.1109/tnn.2002.1000150

Sanjay Goil, Harsha Nagesh, and Alok Choudhary. 1999. 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*. 443–452.

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. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 5439 (1999), 531–537.

Teofilo F Gonzalez. 1985. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science* 38, 2-3 (1985), 293–306.

J. C. Gower and G. J. S. Ross. 1969. Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 18, 1 (1969), 54–64. http://www.jstor.org/stable/2346439

Nizar Grira, Michel Crucianu, and Nozha Boujemaa. 2005. Semi-supervised fuzzy clustering with pairwise-constrained competitive agglomeration. In *Fuzzy Systems, 2005. FUZZ'05. The 14th IEEE International Conference on*. IEEE, 867–872.

Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. 1998. CURE: An efficient clustering algorithm for large databases. *SIGMOD Record* 27, 2 (June 1998), 73–84. DOI:http://dx.doi.org/10.1145/276305.276312

Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. 2000. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems* 25, 5 (2000), 345 – 366. DOI:http://dx.doi.org/10.1016/S0306-4379(00)00022-3

Chetan Gupta and Robert Grossman. 2004. GenIc: A Single Pass Generalized Incremental Algorithm for Clustering. In *Proceedings of the Fourth SIAM International Conference on Data Mining*. SIAM, 147–153.

Valerie Guralnik and George Karypis. 2001. A Scalable Algorithm for Clustering Sequential Data.. In *ICDM* (2008-09-24), Nick Cercone, Tsau Young Lin, and Xindong Wu (Eds.). IEEE Computer Society, 179–186. http://dblp.uni-trier.de/db/conf/icdm/icdm2001.html#GuralnikK01

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Lars Hagen and Andrew B. Kahng. 1992. New Spectral Methods for Ratio Cut Partitioning and Clustering. *IEEE Tansactions on Computer-aided Design* 11, 9 (September 1992), 1074–1085.

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

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

Julia Handl, Joshua Knowles, and Douglas B. Kell. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 15 (2005), 3201–3212. DOI:http://dx.doi.org/10.1093/bioinformatics/bti517

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

J. A. Hartigan. 1975. Clustering algorithms. In *Clustering Algorithms*. John Wiley & Sons, Inc.

J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A k-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28, 1 (1979), 100108.

Alexander Hinneburg, Er Hinneburg, and Daniel A. Keim. 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, Vol. 98. AAAI Press, 58–65.

Alexander Hinneburg and Daniel A. Keim. 1999. Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering.. In *Proceedings of the 25th International Conference on Very Large Data Bases* (2002-01-03), Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie (Eds.). Morgan Kaufmann, 506–517. http://dblp.uni-trier.de/db/conf/vldb/vldb99.html#KeimH99

Ricky Ho. 2012. Machine Learning in R: Clustering. (2012). URL http://horicky.blogspot.com/2012/04/machine-learning-in-r-clustering.html. Accessed Oct. 2015.

Matthew D. Hoffman, David M. Blei, and Francis R. Bach. 2010. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*. 856–864.

Thomas Hofmann and Joachim M. Buhmann. 1997. Pairwise Data Clustering by Deterministic Annealing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19. 1–14. Issue 1.

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

Prodip Hore, Lawrence O Hall, and Dmitry B Goldgof. 2009. A scalable framework for cluster ensembles. *Pattern recognition* 42, 5 (2009), 676–688.

P. V. C. Hough. 1962. Methods and means for recognizing complex patterns. (18 December 1962). US Patent 3069654.

Xiaohua Hu and Illhoi Yoo. 2004. Cluster ensemble and its applications in gene expression analysis. In *Proceedings of the Second Conference on Asia-Pacific Bioinformatics (APBC '04)*, Vol. 29. Australian Computer Society, Inc., Darlinghurst, Australia, 297–302. http://dl.acm.org/citation.cfm?id=976520.976560

Z. Huang. 1998. Extensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 2, 3 (1998), 283–304.

Christopher L. Huntley and Donald E. Brown. 1991. A parallel heuristic for quadratic assignment problems. *Computers & Operations Research* 18, 3 (1991), 275–289. http://dblp.uni-trier.de/db/journals/cor/cor18.html#HuntleyB91

N. Iam-On, T. Boongeon, S. Garrett, and C. Price. 2012. A Link-Based Cluster Ensemble Approach for Categorical Data Clustering. *Knowledge and Data Engineering, IEEE Transactions on* 24, 3 (March 2012), 413–425. DOI:http://dx.doi.org/10.1109/TKDE.2010.268

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

V. Ilango, R. Subramanian, and V. Vasudevan. 2011. Cluster Analysis Research Design model, problems, issues, challenges, trends and tools. *International Journal on Computer Science and Engineering* 3, 8 (2011), 2926–2934.

Marilena V. Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, and others. 2005. MicroRNA gene expression deregulation in human breast cancer. *Cancer Research* 65, 16 (2005), 7065–7070.

A. Jain, R. Duin, and J. Mao. 2000. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22, 1 (2000), 4–37.

Anil K. Jain. 2007. Data Clustering: User's Dilemma. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM '07)*. Springer-Verlag, Berlin, Heidelberg, 1–1. DOI:http://dx.doi.org/10.1007/978-3-540-73499-4_1

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

A. K. Jain and R. C. Dubes. 1988. Algorithms for Clustering Data. In *Prentice-Hall Advanced Reference Series*. Prentice-Hall, Inc.

Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys (CSUR)* 31, 3 (1999), 264–323.

Eshref Januzaj, Hans-Peter Kriegel, and Martin Pfeifle. 2004. DBDC: Density based distributed clustering. In *Advances in Database Technology-EDBT 2004*. Springer, 88–105.

Daxin Jiang, Chun Tang, and Aidong Zhang. 2004. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Trans. on Knowl. and Data Eng.* 16, 11 (Nov. 2004), 1370–1386. DOI:http://dx.doi.org/10.1109/TKDE.2004.68

H. D. Jin. 2002. *Scalable Model-Based Clustering Algorithms for Large Databases and Their Applications*. Ph.D. Dissertation. The Chinese University of Hong Kong.

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

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

G. Karypis, R. Aggarwal, V. Kumar, and Shashi Shekhar. 1999. Multilevel hypergraph partitioning: Applications in VLSI domain. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* 7, 1 (March 1999), 69–79. DOI:http://dx.doi.org/10.1109/92.748202

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

George Karypis and Vipin Kumar. 1998. *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.

L. Kaufman and P. Rousseeuw. 1987. *Clustering by Means of Medoids*. North-Holland.

Leonard Kaufman and Peter J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons.

Paul Kellam, Xiaohui Liu, Nigel Martin, Christine Orengo, Stephen Swift, and Allan Tucker. 2001. Comparing, contrasting and combining clusters in viral gene expression data. In *Proceedings of 6th Workshop on Intelligent Data Analysis in Medicine and Pharmocology*. 56–62.

James F. Kennedy, James Kennedy, and Russell C. Eberhart. 2001. *Swarm Intelligence*. Morgan Kaufmann.

Christopher G. Kevil, Loren Walsh, F. Stephen Laroux, Theodore Kalogeris, Matthew B. Grisham, and J. S. Alexander. 1997. An improved, rapid Northern protocol. *Biochemical and Biophysical Research Communications* 238, 2 (1997), 277–279.

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. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7, 6 (June 2001), 673679. DOI:http://dx.doi.org/10.1038/89044

Eun-Youn Kim, Seon-Young Kim, Daniel Ashlock, and Dougu Nam. 2009. MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics* 10, 1 (2009), 260.

Benjamin King. 1967. Step-wise clustering procedures. *J. Amer. Statist. Assoc.* 62, 317 (1967), 86–101.

D. Klein, S. Kamvar, and C. Manning. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the 19th International Conference on Machine Learning*. 307–314. http://citeseer.ist.psu.edu/klein02from.html

R. W. Klein and R. C. Dubes. 1989. Experiments in projection and clustering by simulated annealing. *Pattern Recognition* 22, 2 (1989), 213–220. http://scholar.google.com.au/scholar.bib?q=info:UOjcuecfod0J:scholar.google.com/&output=citation&hl=en&as_sdt=2000&ct=citation&cd=0

J. Kleinberg. 2002. An impossibility theorem for clustering. *Proceeding Conference Advances in Neural Information Processing Systems* 15 (2002), 463–470.

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, and others. 2012. Comprehensive Molecular Portraits of Human Breast tumours. *Nature* 490, 7418 (2012), 61–70.

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

Jachoon Koo, Yumi Kim, Jeongsik Kim, Miji Yeom, In Chul Lee, and Hong Gil Nam. 2007. A GUS/Luciferase Fusion Reporter for Plant Gene Trapping and for Assay of Promoter Activity with Luciferin-Dependent Control of the Reporter Protein Stability. *Plant and Cell Physiology* 48, 8 (2007), 1121–1131. DOI:http://dx.doi.org/10.1093/pcp/pcm081

Dorota T. Kopycka-Kedzierawski and Ronald J. Billings. 2004. 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 (2004), 201–209.

Hans-Peter Kriegel, Peer Kröger, Matthias Renz, and Sebastian Wurst. 2005. A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data.. In *Data Mining, Fifth IEEE International Conference on* (2005-12-21). IEEE Computer Society, 250–257. http://dblp.uni-trier.de/db/conf/icdm/icdm2005.html#KriegelKRW05

Peer Kröger, Hans-Peter Kriegel, and Karin Kailing. 2004. Density-Connected Subspace Clustering for High-Dimensional Data.. In *SDM* (2004-07-12), Michael W. Berry, Umeshwar Dayal, Chandrika Kamath, and David B. Skillicorn (Eds.), Vol. 4. SIAM. http://dblp.uni-trier.de/db/conf/sdm/sdm2004.html#KroegerKK04

Joseph B. Kruskal. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society* 7, 1 (1956), 48–50.

N. Kumar and R. S. Joshi. 2007. Data Clustering Using Artificial Neural Networks. In *Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007)*. 197–200.

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

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. 2006. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene* 26, 2 (03 July 2006), 312–320. DOI:http://dx.doi.org/10.1038/sj.onc.1209778

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. 2004. 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 (20 Jan. 2004), 811–816. DOI:http://dx.doi.org/10.1073/pnas.0304146101

K. W. Lau, H. Yin, and S. Hubbard. 2006. Kernel self-organising maps for classification. *Neurocomputing* 69, 1618 (2006), 2033 – 2040. DOI:http://dx.doi.org/10.1016/j.neucom.2005.10.003 Selected papers from the 1st International Conference on Brain Inspired Cognitive Systems (BICS 2004).

Laura Lazzeroni and Art Owen. 2000. Plaid Models for Gene Expression Data. *Statistica Sinica* 12, 1 (2000), 61–86.

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

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

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

Li Li, Chang Liu, Fang Wang, Wei Miao, Jie Zhang, Zhiqian Kang, Yihan Chen, and Luying Peng. 2014. Unraveling the Hidden Heterogeneities of Breast Cancer Based on Functional miRNA Cluster. *PLOS ONE* 9, 1 (2014), e87601–v.

Tao Li and Chris H. Q. Ding. 2008. Weighted Consensus Clustering. In *SDM'08*. 798–809.

T. Li, C. Zhang, and M. Ogihara. 2004b. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20, 15 (12 Oct. 2004), 2429–2437.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. ACM, New York, NY, USA, 577–584. DOI:http://dx.doi.org/10.1145/1143844.1143917

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. 2005. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proceedings of the National Academy of Sciences USA* 102, 16 (2005), 5814–5819.

Wei-keng Liao, Ying Liu, and Alok Choudhary. 2004. A grid-based clustering algorithm using adaptive mesh refinement. In *7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining*.

Ricardo Linden. 2009. Clustering Techniques. *Revista de Sistemas de Informação da FSMA* 4 (2009), 18–36.

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

Jingwei Liu and Meizhi Xu. 2008. Kernelized fuzzy attribute C-means clustering algorithm. *Fuzzy Sets and Systems* 159, 18 (2008), 2428–2445. http://dblp.uni-trier.de/db/journals/fss/fss159.html#LiuX08

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

Eric F. Lock and David B. Dunson. 2013. Bayesian consensus clustering. *Bioinformatics* 29, 20 (2013), 2610–2616. DOI:http://dx.doi.org/10.1093/bioinformatics/btt425

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, and others. 2005. MicroRNA expression profiles classify human cancers. *Nature* 435, 7043 (2005), 834–838.

Huilan Luo, Furong Jing, and Xiaobing Xie. 2006. Combining Multiple Clusterings using Information Theory based Genetic Algorithm. In *Computational Intelligence and Security, 2006 International Conference on*, Vol. 1. 84–89. DOI:http://dx.doi.org/10.1109/ICCIAS.2006.294095

P. C. H. Ma, K. C. C. Chan, Xin Yao, and D. K. Y. Chiu. 2006. An evolutionary clustering algorithm for gene expression microarray data analysis. *Evolutionary Computation, IEEE Transactions on* 10, 3 (june 2006), 296 – 314. DOI:http://dx.doi.org/10.1109/TEVC.2005.859371

J. B. MacQueen. 1967. 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 (1967), 281297.

Sara C. Madeira and Arlindo L. Oliveira. 2004. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1, 1 (2004), 24–45. DOI:http://dx.doi.org/10.1109/TCBB.2004.2

S. W. Mahfoud. 1995. *Niching Methods for Genetic Algorithms*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

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. 1999. Vascular Channel Formation by Human

Melanoma Cells *in Vivo* and *in Vitro*: Vasculogenic Mimicry. *The American Journal of Pathology* 155, 3 (1999), 739–752.

Jianchang Mao and Anil K. Jain. 1996. A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks* 7, 1 (1996), 16–29. http://dblp.uni-trier.de/db/journals/tnn/tnn7.html#MaoJ96

Thomas M. Martinetz and Klaus J. Schulten. 1991. A "Neural Gas" Network Learns Topologies. In *Proceedings of the International Conference on Artificial Neural Networks* (Espoo, Finland), Teuvo Kohonen, Kai Mäkisara, Olli Simula, and Jari Kangas (Eds.). Amsterdam; New York: North-Holland, 397–402.

Frederic H. Martini. 2012. *e-Study Guide for Fundamentals of Anatomy & Physiology*. Cram101 Textbook Reviews. https://books.google.com/books?id=E54j4oEp3RsC

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

G. J. McLachlan and K. E. Basford. 1988. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc, New York / Basel.

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

Geoffrey J. McLachlan, Kim-Anh Do, and Christophe Ambroise. 2005. *Microarrays in Gene Expression Studies*. John Wiley & Sons, Inc., 1–29. DOI: http://dx.doi.org/10.1002/047172842X.ch1

Geoffrey J. McLachlan, Shu-Kay Ng, and Richard Bean. 2006. Robust cluster analysis via mixture models. *Austrian Journal of Statistics* 35, 2 (2006), 157–174.

Marina Meila and Jianbo Shi. 2001. Learning Segmentation by Random Walks. In *In Advances in Neural Information Processing Systems*. MIT Press, 873–879.

Markus Metzler, Monika Wilda, Kerstin Busch, Susanne Viehmann, and Arndt Borkhardt. 2004. High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes, Chromosomes and Cancer* 39, 2 (2004), 167–169.

Michael Z Michael, Susan M O'Connor, Nicholas G van Holst Pellekaan, Graeme P Young, and Robert J James. 2003. Reduced Accumulation of Specific MicroRNAs in Colorectal Neoplasia. *Molecular Cancer Research* 1, 12 (2003), 882–891.

Boriana L. Milenova and Marcos M. Campos. 2002. O-cluster: Scalable clustering of large high dimensional data sets. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 290–297.

Gabriela Moise, Jörg Sander, and Martin Ester. 2006. P3C: A Robust Projected Clustering Algorithm. In *ICDM* (2007-01-09). IEEE Computer Society, 414–425. http://dblp.uni-trier.de/db/conf/icdm/icdm2006.html#MoiseSE06

Gabriela Moise, Jörg Sander, and Martin Ester. 2008. Robust projected clustering. *Knowledge and Information Systems* 14, 3 (2008), 273–298. http://dblp.uni-trier.de/db/journals/kais/kais14.html#MoiseSE08

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

B. Moore. 1989. ART1 and pattern clustering. *Proceedings Connectionist Models Summer School* (1989), 174–185.

Tadeusz Morzy, Marek Wojciechowski, and Maciej Zakrzewicz. 1999. Pattern-Oriented Hierarchical Clustering. In *Proceedings of the third East-European Symposium on Advances in Databases and Information Systems ADBIS99, Slovenia, LNCS 1691*. 179–190.

T. S. Motzkin and E. G. Straus. 1965. Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics* 17, 4 (1965), 533–540. http://math.ca/10.4153/CJM-1965-053-6#

Anirban Mukhopadhyay and Ujjwal Maulik. 2011. A multiobjective approach to MR brain image segmentation. *Applied Soft Computing* 11, 1 (2011), 872–880. http://dblp.uni-trier.de/db/journals/asc/asc11.html#MukhopadhyayM11

U. R. Müller and D. V. Nicolau. 2005. *Microarray Technology and Its Applications*. Springer. http://books.google.com/books?id=-2DRvljktw0C

Fionn Murtagh. 1983. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* 26, 4 (1983), 354–359.

Megha Nangia. 2012. Partitional Clustering. *ACM student chapter, SIGKDD Presentation* (February 2012).

NCI. 2013a. A Snapshot of Leukemia. (2013). URL http://www.cancer.gov/researchandfunding/snapshots/leukemia. Accessed Oct. 2013.

NCI. 2013b. Understanding Cancer Series. (2013). URL http://www.cancer.gov/cancertopics/understandingcancer. Accessed Oct. 2013.

S. B. Needleman and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48, 3 (March 1970), 443–453. http://view.ncbi.nlm.nih.gov/pubmed/5420325

Jaroslav Nešetřil, Eva Milková, and Helena Nešetřilová. 2001. Otakar Boruvka on minimum spanning tree problem translation of both the 1926 papers, comments, history. *Discrete Mathematics* 233, 1 (2001), 3–36.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*. MIT Press, 849–856.

Raymond T. Ng and Jiawei Han. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 144–155. http://dl.acm.org/citation.cfm?id=645920.672827

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

Tim Oates, Laura Firoiu, and Paul R. Cohen. 2001. Using Dynamic Time Warping to Bootstrap HMM-Based Clustering of Time Series.. In *Sequence Learning* (2002-01-03) *(Lecture Notes in Computer Science)*, Ron Sun and C. Lee Giles (Eds.), Vol. 1828. Springer, 35–52. http://dblp.uni-trier.de/db/conf/seqlearn/seqlearn2001.html#OatesFC01

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. 2011. miRNA Expression in Colon Polyps Provides Evidence for a Multihit Model of Colon Cancer. *PLoS ONE* 6, 6 (06 2011), e20465–e20465. DOI:http://dx.doi.org/10.1371/journal.pone.0020465

Elizabeth O'Day and Ashish Lal. 2010. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Research* 12, 2 (2010), 201–201. DOI:http://dx.doi.org/10.1186/bcr2484

Atsuyuki Okabe and Kokichi Sugihara. 2012. *Spatial Analysis Along Networks: Statistical and Computational Methods*. John Wiley & Sons.

Niina Päivinen. 2005. Clustering with a minimum spanning tree of scale-free-like structure. *Pattern Recognition Letters* 26, 7 (2005), 921–930.

Linus Pauling, A. Harvey Itano, S. J. Singer, and Ibert C. Wells. 1949. Sickle Cell Anemia, a Molecular Disease. *Science* 110, 2865 (1949), 543–548.

Massimiliano Pavan and Marcello Pelillo. 2007. Dominant Sets and Pairwise Clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29, 1 (Jan. 2007), 167–172. DOI:http://dx.doi.org/10.1109/TPAMI.2007.10

Z. Pawlak. 1991. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Springer Netherlands. http://books.google.com/books?id=MJPLCqIniGsC

Karl Pearson. 1901. 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 (1901), 559–572.

W. R. Pearson and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci. U. S. A.* 85, 8 (1 April 1988), 2444–2448. DOI:http://dx.doi.org/10.1073/pnas.85.8.2444

Dan Pelleg and Andrew Moore. 1999. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 277–281.

Dan Pelleg and Andrew Moore. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Pattern Recognition Letters, of: Proceedings of the Seventeenth International Conference on Machine Learning*. 727–734.

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, and others. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences* 96, 16 (1999), 9212–9217.

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. 2000. Molecular portraits of human breast tumours. *Nature* 406, 6797 (17 Aug. 2000), 747–752. DOI:http://dx.doi.org/10.1038/35021093

Piercenet. 2013. Overview of Immunohistochemistry. (2013). URL http://www.piercenet.com/method/overview-immunohistochemistry. Accessed Oct. 2013.

Harun Pirim, Burak Ekşioğlu, Andy D. Perkins, and Çetin Yüceer. 2012. Clustering of high throughput gene expression data. *Computers & Operations Research* 39, 12 (2012), 3046–3061.

Harun Pirim, Dilip Gautam, Tanmay Bhowmik, Andy D. Perkins, and Burak Ekioglu. 2011. Performance of an Ensemble Clustering Algorithm on Biological Data Sets. *Mathematical and Computational Applications* 16, 1 (2011), 87–96.

Clara Pizzuti and Domenico Talia. 2003. P-AutoClass: Scalable Parallel Clustering for Mining Large Data Sets. *Knowledge and Data Engineering, IEEE Transactions on* 15, 3 (March 2003), 629–641. DOI:http://dx.doi.org/10.1109/TKDE.2003.1198395

J. Platt. 1999. *Advances in Kernel Methods: Support Vector Learning*. MIT press, Cambridge, MA, Chapter Fast training of SVMs using sequential minimal optimization, 185–208.

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

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

R. C. Prim. 1957. Shortest connection networks and some generalizations. *Bell System Technology Journal* 36, 6 (1957), 1389–1401.

Nobel Prize. 2013. Life and Discoveries of Camillo Golgi. (2013). URL http://www.nobelprize.org/nobel_prizes/medicine/laureates/1906/golgi-article.html. Accessed Oct. 2013.

Cecilia Magdalena Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. 2002. A Monte Carlo algorithm for fast projective clustering.. In *SIGMOD Conference* (2003-02-06), Michael J. Franklin, Bongki Moon, and Anastassia Ailamaki (Eds.). ACM, 418–427. http://dblp.uni-trier.de/db/conf/sigmod/sigmod2002.html#ProcopiucJAM02

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

Sridhar Ramaswamy and Todd R. Golub. 2002. DNA Microarrays in Clinical Oncology. *Journal of Clinical Oncology* 20, 7 (2002), 1932–1941. http://jco.ascopubs.org/cgi/content/abstract/20/7/1932

Marco Ramoni, Paola Sebastiani, and Paul R. Cohen. 2002. Bayesian Clustering by Dynamics. *Machine Learning* 47, 1 (2002), 91–121. http://dblp.uni-trier.de/db/journals/ml/ml47.html#RamoniSC02

John I. Risinger, G. Larry Maxwell, G. V. Chandramouli, Amir Jazaeri, Olga Aprelikova, Tricia Patterson, Andrew Berchuck, and J. Carl Barrett. 2003. Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Research* 63, 1 (1 Jan. 2003), 6–11. http://view.ncbi.nlm.nih.gov/pubmed/12517768

Jon E Roeckelein. 1998. *Dictionary of Theories, Laws, and Concepts in Psychology*. Greenwood Publishing Group.

Roco Romero-Zliz, Cristina Rubio-Escudero, J. P. Cobb, Francisco Herrera, Oscar Cordn, and Igor Zwir. 2008. 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 (2008), 679–701. http://dblp.uni-trier.de/db/journals/tec/tec12.html#Romero-ZalizRCHCZ08

Carlos Ruiz, Myra Spiliopoulou, and Ernestina Menasalvas. 2007. C-DBSCAN: Density-based clustering with constraints. In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*. Springer, 216–223.

Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery* 2, 2 (June 1998), 169–194. DOI:http://dx.doi.org/10.1023/A:1009745219419

Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 5235 (1995), 467–470.

E. Schikuta. 1996. Grid-clustering: An efficient hierarchical clustering method for very large data sets. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on* 2 (Aug. 1996), 101–105. DOI:http://dx.doi.org/10.1109/ICPR.1996.546732

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

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

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 10, 5 (1 July 1998), 1299–1319. DOI:http://dx.doi.org/10.1162/089976698300017467

A. J. Scott and Michael J Symons. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* (1971), 387–397.

Shokri Z. Selim and K. Alsultan. 1991. A simulated annealing algorithm for the clustering problem. *Pattern Recognition* 24, 10 (1991), 1003–1008. http://dblp.uni-trier.de/db/journals/pr/pr24.html#SelimA91

Gholamhosein Sheikholeslami, Surojit Chatterjee, and Aidong Zhang. 1998. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In *Proceedings of the 24rd International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., 428–439.

Weiguo Sheng, Allan Tucker, and Xiaohui Liu. 2004. *Clustering with Niching Genetic K-Means Algorithm*. Springer Berlin / Heidelberg, 162–173. http://www.springerlink.com/content/bgtg1qb9c8x54qay/?p=202339f8b2d741a495b549b2b7e4bda3&pi=0

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *IEEE. Reprinted from IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (August 2000), 888–905. http://dx.doi.org/10.1109/34.868688

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

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. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 2 (March 2002), 203–209. http://view.ncbi.nlm.nih.gov/pubmed/12086878

Donna K. Slonim. 2002. From patterns to pathways: Gene expression data analysis comes of age. *Nature Genetics* 32 Suppl (01 Dec. 2002), 502–508. DOI:http://dx.doi.org/10.1038/ng1033

Donna K. Slonim, Pablo Tamayo, Jill P. Mesirov, Todd R. Golub, and Eric S. Lander. 2000. Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*. ACM, 263–272.

Slowmo. 2010. Example 2D space with subspace clusters. (2010). URL http://commons.wikimedia.org/wiki/File:SubspaceClustering.png. Accessed Oct. 2015.

Temple F. Smith and Michael S. Waterman. 1980. New Stratigraphic Correlation Techniques. *The Journal of Geology* 88, 4 (Jul. 1980), 451–457.

Padhraic Smyth. 1996. Clustering Sequences with Hidden Markov Models. In *Advances in Neural Information Processing* (2003-05-28), Michael Mozer, Michael I. Jordan, and Thomas Petsche (Eds.). MIT Press, 648–654. http://dblp.uni-trier.de/db/conf/nips/nipsN1996.html#Smyth96

Padhraic Smyth. 1999. Probabilistic Model-Based Clustering of Multivariate and Sequential Data. In *In Proceedings of Artificial Intelligence and Statistics*. Morgan Kaufmann, 299–304.

P. Sneath and R. Sokal. 1973. Numerical Taxonomy. In *Numerical Taxonomy*. W. H. Freeman and Company.

Peter H. A. Sneath. 1957. The application of computers to taxonomy. *Journal of general microbiology* 17, 1 (1957), 201–226.

The Leukemia & Lymphoma Society. 2013. ALL Subtypes. (2013). URL http://www.lls.org/diseaseinformation/leukemia/acutelymphoblasticleukemia/allsubtypes/. Accessed Oct. 2013.

R. R. Sokal and C. D. Michener. 1958. A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Scientific Bulletin* 38 (1958), 1409–1438.

Y. Song, W. Wang, X. Qu, and S. Sun. 2009. Effects of hypoxia inducible factor-1alpha (HIF-1alpha) on the growth & adhesion in tongue squamous cell carcinoma cells. *The Indian Journal of Medical Research* 129, 2 (February 2009), 154–163.

T. Sorensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons. *Biologiske Skrifter* 5 (1948), 1–34.

H. Spath. 1980. Cluster Analysis Algorithms. In *Cluster Analysis Algorithms*. West Sussex, Ellis Horwood Limited.

Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21, 5 (March 2005), 631–643. DOI:http://dx.doi.org/10.1093/bioinformatics/bti033

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. *Efficient algorithms for creating product catalogs*. Technical Report. DTIC Document.

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

Alexander Strehl and Joydeep Ghosh. 2002. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research* 3 (2002), 583–617. http://dblp.uni-trier.de/db/journals/jmlr/jmlr3.html#StrehlG02

Ron Sun and C. Lee Giles. 2001. *Sequence Learning: Paradigms, Algorithms, and Applications*. Vol. 1828. Springer.

Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow, and Donald Geman. 2005. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21, 20 (2005), 3896–3904. DOI:http://dx.doi.org/10.1093/bioinformatics/bti631

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

Chun Tang and Aidong Zhang. 2002. An iterative strategy for pattern discovery in high-dimensional data sets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*. ACM, 10–17.

Chun Tang, Li Zhang, Aidong Zhang, and Murali Ramanathan. 2001. 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*. 41–48. http://dblp.uni-trier.de/db/conf/bibe/bibe2001.html#TangZZR01

Sean Taylor, Michael Wakem, Greg Dijkman, Marwan Alsarraj, and Marie Nguyen. 2010. A practical approach to RT-qPCR – Publishing data that conform to the MIQE guidelines. *Methods* 50, 4 (2010), S1 – S5. DOI:http://dx.doi.org/10.1016/j.ymeth.2010.01.005

S. Theodoridis and K. Koutroumbas. 2006. *Pattern Recognition, Third Edition*. Academic Press, Inc., Orlando, FL, USA.

S. C. A. Thomopoulos, D. K. Bougoulias, and Chin-Der Wann. 1995. Dignet: An unsupervised-learning clustering algorithm for clustering and data fusion. *Aerospace and Electronic Systems, IEEE Transactions on* 31, 1 (Jan 1995), 21–38. DOI:http://dx.doi.org/10.1109/7.366289

Alexander P. Topchy, Anil K. Jain, and William F. Punch. 2003. Combining Multiple Weak Clusterings.. In *Proceedings of the IEEE International Conference on Data Mining* (2004-12-01). IEEE Computer Society, 331–338. http://dblp.uni-trier.de/db/conf/icdm/icdm2003.html#TopchyJP03

Alexander P. Topchy, Anil K. Jain, and William F. Punch. 2004. A Mixture Model for Clustering Ensembles.. In *Proceedings SIAM International Conference on Data Mining*. SIAM.

Laura J. Van't Veer and René Bernards. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452, 7187 (3 April 2008), 564–570. DOI:http://dx.doi.org/10.1038/nature06915

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

Victor E. Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W. Kinzler, and others. 1995. Serial analysis of gene expression. *Science* 270, 5235 (1995), 484–487.

Boaz Vigdor and Boaz Lerner. 2007. The Bayesian ARTMAP. *IEEE Transactions on Neural Networks* 18, 6 (2007), 1628–1644. http://dblp.uni-trier.de/db/journals/tnn/tnn18.html#VigdorL07

Michail Vlachos, Jessica Lin, Eamonn Keogh, and Dimitrios Gunopulos. 2003. A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series. In *In Proceedings Workshop on Clustering High Dimensionality Data and Its Applications*. 23–30.

K. Wagstaff and C. Cardie. 2000. Clustering with Instance-level Constraints. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)* (2000), 1103–1110.

Hongjun Wang, Hanhuai Shan, and Arindam Banerjee. 2011. Bayesian cluster ensembles. *Statistical Analysis and Data Mining* 4, 1 (2011), 54–70. DOI:http://dx.doi.org/10.1002/sam.10098

Haiying Wang, Huiru Zheng, and Francisco Azuaje. 2007. 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 (2007), 163–175. DOI:http://dx.doi.org/10.1109/TCBB.2007.070204

Wei Wang, Jiong Yang, and Richard R. Muntz. 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld (Eds.). Morgan Kaufmann, 186–195.

Wei Wang, Jiong Yang, and Richard R. Muntz. 1999. STING+: An Approach to Active Spatial Data Mining.. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, Masaru Kitsuregawa, Michael P. Papazoglou, and Calton Pu (Eds.). IEEE Computer Society, 116–125. http://dblp.uni-trier.de/db/conf/icde/icde99.html#WangYM99

Xiang Wang and Ian Davidson. 2010. Active spectral clustering. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 561–568.

Chin-Der Wann and Stelios C. A. Thomopoulos. 1997. A Comparative Study of Self-organizing Clustering Algorithms Dignet and ART2. *Neural Networks* 10, 4 (1997), 737–753. http://dblp.uni-trier.de/db/journals/nn/nn10.html#WannT97

Joe H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* 58, 301 (1963), 236–244.

Max Welling. 2005. Learning in Markov random fields with contrastive free energies. In *In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. 397–404.

O. Wildi. 2010. *Data Analysis in Vegetation Ecology*. John Wiley & Sons. http://books.google.com/books?id=CIoYrGrlABcC

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

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, and others. 2010. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research* 16, 19 (2010), 4864–4875.

James R. Williamson. 1996. Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multidimensional Maps. *Neural Networks* 9, 5 (1996), 881–897. http://dblp.uni-trier.de/db/journals/nn/nn9.html#Williamson96

John H. Wolfe. 1970. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 5, 3 (1970), 329–350.

Donald C. Wunsch, Thomas P. Caudell, C. David Capps, Robert J. Marks, and R. Aaron Falk. 1993. An optoelectronic implementation of the adaptive resonance neural network. *IEEE Transactions on Neural Networks* 4, 4 (1993), 673–684. http://dblp.uni-trier.de/db/journals/tnn/tnn4.html#WunschCCMF93

Chen Xiaoyun, Chen Yi, Qi Xiaoli, Yue Min, and He Yanshan. 2009. PGMCLU: A novel parallel grid-based clustering algorithm for multi-density datasets. In *Web Society, 2009. SWS'09. 1st IEEE Symposium on*. IEEE, 166–171.

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

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2003. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems* (2003), 521–528.

Yimin Xiong and Dit-Yan Yeung. 2004. Time series clustering with ARMA mixtures. *Pattern Recognition* 37, 8 (2004), 1675–1689. http://dblp.uni-trier.de/db/journals/pr/pr37.html#XiongY04

Maoxiong Xu. 2011. *A HMM Approach to Identifying Distinct DNA Methylation Patterns for Subtypes of Breast Cancers*. Ph.D. Dissertation. The Ohio State University.

R. Xu, G. C. Anagnostopoulos, and D. C. Wunsch. 2007. 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 (Jan-Mar 2007), 65–77.

Rui Xu and D. Wunsch. 2005. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on* 16, 3 (may 2005), 645 –678. DOI:http://dx.doi.org/10.1109/TNN.2005.845141

Rui Xu and D. Wunsch. 2009. *Clustering*. IEEE/Wiley.

Rui Xu and D. C. Wunsch. 2010. Clustering Algorithms in Biomedical Research: A Review. *Biomedical Engineering, IEEE Reviews in* 3 (2010), 120.

Sen Xu, Zhimao Lu, and Guochang Gu. 2008. An Efficient Spectral Method for Document Cluster Ensemble. In *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*. 808 –813. DOI:http://dx.doi.org/10.1109/ICYCS.2008.228

Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jörg Sander. 1998. A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. In *Proceedings of the Fourteenth International Conference on Data Engineering (ICDE '98)*. IEEE Computer Society, Washington, DC, USA, 324–331. http://dl.acm.org/citation.cfm?id=645483.653621

Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas A. J. Schweiger. 2007. SCAN: A structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 824–833.

Ronald R. Yager. 2000. Intelligent control of the hierarchical agglomerative clustering process. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 30, 6 (2000), 835–845. http://dblp.uni-trier.de/db/journals/tsmc/tsmcb30.html#Yager00b

Donghui Yan, Ling Huang, and Michael I Jordan. 2009. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 907–916.

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

Zhao Yanchang and Song Junde. 2001. GDILC: A grid-based density-isoline clustering algorithm. In *Info-tech and Info-net, 2001. Proceedings. ICII 2001 - Beijing. 2001 International Conferences on*, Vol. 3. 140–145. DOI:http://dx.doi.org/10.1109/ICII.2001.983048

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

Qiang Yang and Xindong Wu. 2006. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology and Decision Making (IJITDM)* 05, 04 (2006), 597–604. http://EconPapers.repec.org/RePEc:wsi:ijitdm:v:05:y:2006:i:04:p:597-604

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. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell* 1, 2 (March 2002), 133–143. http://view.ncbi.nlm.nih.gov/pubmed/12086872

James Yolkowski. 2014. The Clustering Illusion. (2014). URL http://mathlair.allfunandgames.ca/clustering.php. Accessed Oct. 2015.

Stella X. Yu and Jianbo Shi. 2003. Multiclass Spectral Clustering. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2 (ICCV '03)*. IEEE Computer Society, Washington, DC, USA, 313–319. http://dl.acm.org/citation.cfm?id=946247.946658

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

Stefanos Zafeiriou and Nikolaos A. Laskaris. 2008. On the Improvement of Support Vector Techniques for Clustering by Means of Whitening Transform. *Signal Processing Letters, IEEE* 15 (2008), 198–201. http://dblp.uni-trier.de/db/journals/spl/spl15.html#ZafeiriouL08

Charles T Zahn. 1971. Graph-theoretical methods for detecting and describing gestalt clusters. *Computers, IEEE Transactions on* 100, 1 (1971), 68–86.

B. Zhang. 2001. Generalized k-harmonic means – Dynamic weighting of data in un-supervised learning. *Proceedings of the 1st SIAM ICDM, Chicago, IL, USA* (2001), 1–13.

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

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An efficient data clustering method for very large databases. *SIGMOD Record* 25, 2 (June 1996), 103–114. DOI:http://dx.doi.org/10.1145/235968.233324

Chuan Zhou. 2003. *A Bayesian Model for Curve Clustering with Application to Gene Expression Data Analysis*. Ph.D. Dissertation. University of Washington.

Shangming Zhou and John Q. Gan. 2004. An Unsupervised Kernel Based Fuzzy C-means Clustering Algorithm with Kernel Normalisation. *International Journal of Computational Intelligence and Applications* 04, 04 (2004), 355–373. DOI:http://dx.doi.org/10.1142/S1469026804001379

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

Received May 1111; revised May 1111; accepted May 1111