Stony Brook University

New Development in Cluster Analysis and Other Related Multivariate Analysis Methods

A Dissertation Presented

by

Shaonan Zhang

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2011

Stony Brook University

The Graduate School

Shaonan Zhang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Wei Zhu

Dissertation Advisor

Professor, Deputy Chair, Department of Applied Mathematics and Statistics

Jiaqiao Hu

Chairperson of Defense

Assistant Professor, Department of Applied Mathematics and Statistics

Haipeng Xing

Member

Assistant Professor, Department of Applied Mathematics and Statistics

Helene D. Benveniste

Outside Member

Professor, Vice Chair for Research, Department of Anesthesiology

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

New Development in Cluster Analysis and Other Related Multivariate Analysis Methods

By

Shaonan Zhang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2011

Cluster analysis is a multivariate analysis method aimed at (1) unraveling the natural groupings embedded within the data, and (2) dimension reduction. With the wide application of cluster analysis in the diversified modern research/business fields including machine learning, bioinformatics, medical image analysis, pattern recognition, market research and global climate research, many clustering algorithms have been developed to date. However, novel and/or special circumstances always call for better customized cluster analysis methods, and thus this thesis.

This thesis work consists of two parts. In the first part, we extend the modern multiple-objective cluster analysis from using a single set of features to multiple distinct sets of features by developing the novel compound clustering method and the constrained clustering method. We also developed a new statistic, the "complete linkage" R^2 along with the well-known largest average silhouette, to determine the optimal number of clusters in the compound clustering. The novel compound/constrained clustering methods are illustrated through a gene microarray study with both gene expression data and gene function information.

In the second part of this thesis we propose a novel algorithm for the weighted k-means clustering. Weighted k-means clustering is an extension of the k-means clustering in which a set of nonnegative weights are assigned to all the variables. We first derived the optimal variable weights for weighted k-means clustering in order to obtain more meaningful and interpretable clusters. We then improved the current weighted k-means clustering method (Huh and Lim 2009) by incorporating our novel algorithm to obtain global-optimal guaranteed variable weights based on the method of Lagrange multiplier and the Karush-Kuhn-Tucker conditions. Here we first present the related theoretical formulation and derivation of the optimal weights. Then we provide an iteration-based computing algorithm to calculate such optimal weights. Numerical examples on both simulated and well known real data are provided to illustrate our method. It is shown that our method outperforms the original proposed method in terms of classification accuracy, stability and computation efficiency.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation work would not have been possible without the support of many people.

First and foremost I would like to offer my sincerest gratitude to my advisor, Professor Wei Zhu, who has been supporting me during my graduate study in Stony Brook University with her patience and knowledge in the past four and a half years. She is not only a great advisor leading me into the statistical world, but also an excellent statistician teaching me everything I know about statistics. Without her excellent guidance and encouragement, this dissertation would not have been completed or written. Furthermore, outside the research world, she is just like a wonderful, nice and sweet mom. I felt so lucky to have her as my advisor. One just simply could not wish to find a better or nicer advisor.

I also would like to thank Dr. Helene Benveniste and Dr. Hedok Lee. It is an honor for me to work with them in the past two years. I enjoyed our fruitful collaborations in Brookhaven National Laboratory and Stony Brook Medical Center. Specially, I want to express my grateful gratitude to Dr. Benveniste, who has been supportive to me all the time financially and mentally. Although she is not a statistician, she has taught me how to conduct scientific research, how to efficiently communicate with others and how to be professional.

Special thanks go to Dr. Haipeng Xing and Dr. Jiaqiao Hu for serving on my defense committee. I also would like to thank Dr. Hu for his valuable suggestions and help in developing the second part of my dissertation.

I am indebted to many of my colleagues for supporting me in my research, to Dr. Tianyi Zhang for frequent discussion in developing the first part of my dissertation; to Dr. Xiao Wu and Hongyan Chen for helping in the biological interpretation of my results; to Shirley Leong for helping me finalizing this dissertation work; to all my other colleagues for providing an excellent atmosphere for doing research in our research group.

Last but not least, I owe my deepest gratitude to my parents for their care, love and support. I also would like to thank my girlfriend, Huan Qi. She was always there cheering me up and stood by me through the good and bad times.

This dissertation is dedicated to all of them.

Chapter 1

Introduction

1.1 Overview

Born to solving quantitative problems arising from all research and business disciplines, statistics is facing increasing challenges from the emerging applications and problems from the rapidly evolving science and industry fields. Previously, statistical problems often came from agricultural and industrial experiments with limited size and scope. In the recent decades, however, people are able to gather huge amount of data that are not only large in sample size, but often more so in dimension – i.e. the number of variables; with the explosive growth of computer and information technology, it becomes more feasible and attractive to let the machine to discover the hidden patterns and useful information from the data, and for dimension reduction -- the so-called "(automated) learning from the data", or "machine learning".

One important feature of machine learning is that the patterns and information are not usually suggested by experts in the field, but rather, extracted and optimized by data-driven techniques. Rising up to the challenge, a large number of statistical approaches have been developed and widely applied to machine learning tasks such as pattern recognition (Nakatani and Hirschberg 1993; Breiman 2001), data mining (Mitchell 1999; Eyke 2005), bioinformatics (Vlahou, et al. 2003; Izmirlian 2004; Larrañaga, et al. 2006), medical image analysis (Pham, et al. 2000; Igor 2001; Rahman, et al. 2007), natural language processing (Ratnaparkhi 1999; Collobert and Weston 2008), document classification and credit scoring (Grossman and Poor 1996; Huang, et al. 2007a). Based on whether the sample is labeled, machine learning algorithms can be classified as supervised and unsupervised learning (Bernd

1994; Figueiredo and Jain 2002; Cohen, et al. 2004; Caruana and Niculescu-Mizil 2006; Chen, et al. 2009; Hastie, et al. 2009).

Supervised learning studies the objects with “label”, which is an outcome measurement of interest. The outcome could be a categorical variable (classes) in classification-type problems or a continuous variable (predictor) in regression-type problems. However, in most cases, supervised machine learning refers to the classification-type problems. Thus a typical supervised learning problem is to predict the outcome measurement (say, “diseased” or “normal”) based on a set of features or attributes (such as the expression levels of a set of biomarkers). A general process of applying such supervised learning to a real-world problem has been described by Kotsiantis (2007). Supervised learning can be very useful for diagnostics/predictions. Consequently, many supervised learning methods have been developed and continuously improved to date including Decision Tree (Kass 1980; Breiman, et al. 1984; Quinlan 1986), Logistic Regression (Menard 2001; Zhu and Hastie 2004), Nearest Neighbor (Dasarathy 1991; Boiman, et al. 2008), Discriminant Analysis (Press and Wilson 1978; McLachlan 2005), Neural Network (Ripley 1994; Bishop 1995; Ripley 1996), Bayesian Network (Jensen 1996; Friedman and Koller 2003; Zou and Conzen 2005), Support Vector Machine (Burges 1998; Suykens and Vandewalle 1999) and Random Forest (Breiman 2001). A thorough discussion on the recent development in supervised learning can be found in the review by Kotsiantis (2007) as well.

However, real world is full of mystery and we are often confronted with realms and phenomena where human beings have never known nor explored before. When it is difficult, too expensive, or simply impossible, to label a sample with it true class, unsupervised learning algorithm is then extremely useful to explore and reveal the hidden data structure based on unlabeled objects. Another situation call for unsupervised learning is “dimension reduction” when one faces a data set with more variables (dimension) than the number of subjects (sample size). In this case, one finds it virtually impossible to apply the usual statistical analyses without first reducing the dimension of the data. The central problem in unsupervised learning is to find natural groupings, or clusters, in multidimensional data, based on measured or perceived

similarities, which can be obtained through a set of features, among the objects (Jain and Dubes (1988). Compared to supervised learning, unsupervised learning can be more challenging due to the lack of label information. Without the “ground truth” information, it is very difficult to evaluate the model and make the adjustment. Major unsupervised learning approaches include feature extraction techniques (e.g. Principal Component Analysis (PCA) (Wold, et al. 1987; Schölkopf, et al. 1997), Multidimensional Scale(MDS) (Green and Carmone 1969; Reidenbach and Robin 1990), Self-Organized Map (Kohonen 1990; Michael 2000; Vesanto and Alhoniemi 2000)), and Cluster Analysis. To certain extent, feature extraction methods can also be considered as clustering algorithms. Ding and He (Ding and He 2004), for example, had discussed the connection and relationship between k-means clustering and PCA in their 2004 paper.

Cluster analysis is a very important and versatile unsupervised learning technique that has seen applications in a wide range of fields such as data mining (Judd, et al. 1998), information retrieval (Carpineto and Romano 1996; Bhatia and Deogun 1998; Messai, et al. 2008), image segmentation (Frigui and Krishnapuram 1999; Tung, et al. 2010), and bioinformatics (Eisen, et al. 1998; Andreopoulos, et al. 2009). This thesis is devoted solely to the improvement and generalization of traditional cluster analysis methods as summarized in the following section.

1.2 Contributions

The focus of this thesis, is on the theory and application of cluster analysis, both hierarchical clustering and partitional clustering where we propose a generalized method for hierarchical clustering and an improved algorithm for partitional clustering analysis. Validation studies confirmed the increased versatility and efficiency of our methods.

1.2.1 Multi-objective Hierarchical Clustering

Traditional cluster analysis is data-driven algorithms without prior information. It has been shown in some cases, incorporating knowledge from multiple sources with

multiple objectives could lead to enhanced performance of existing clustering methods (Cheng, et al. 2004; Huang and Pan 2006). We extend the modern multiple-objective cluster analysis from using a single set of features to multiple distinct sets of features by developing the novel compound clustering method and the constrained clustering method. We have also developed a new statistic, the "complete linkage" R^2 along with the well-known largest average silhouette, to determine the optimal number of clusters in the compound clustering. The novel compound/constrained clustering methods are illustrated through a gene microarray study with both gene expression data and gene function information.

1.2.2 Weighted K-means Clustering

K-means clustering, a partitional clustering algorithm, is widely used because it is easy to implement and interpret. Weighted k-means clustering is an extension of the traditional k-means clustering in which a set of nonnegative weights, possibly unequal, are assigned to all the variables. Solid improvement on clustering performance has been reported by assigning heterogeneous variable weights when performing k-means clustering (Tseng 2007; Shen, et al. 2010). In this thesis, we improve the current weighted k-means clustering method (Huh and Lim 2009) in two aspects. First, we derive the global-optimal guaranteed variable weights for weighted k-means clustering theoretically utilizing the method of Lagrange multiplier and the Karush-Kuhn-Tucker conditions. Subsequently, we improve the current weighted k-means clustering method by incorporating our novel algorithm to obtain global-optimal guaranteed variable weights based on the method of Lagrange multiplier and the Karush-Kuhn-Tucker conditions. Numerical examples on both simulated and well known real data are provided to illustrate our method. It is shown that our method outperforms the original weighted K-means clustering method in terms of classification accuracy, stability and computation efficiency.

1.3 Thesis Structure and Overview

This thesis work is organized as follows. In Chapter 2, we provide a broad literature review of cluster analysis. Chapter 3 and Chapter 4 are devoted to of the novel multi-objective hierarchical clustering method. In Chapter 3, background and several existing multi-objective clustering methods are introduced. In Chapter 4, we describe our proposed compound/constrained clustering methods with application to a temporal gene microarray study. Starting from Chapter 5, we move on to the second contribution of this thesis – namely, the improved weighted k-means algorithm. Chapter 5 provides a literature review on k-means clustering and motivates the modern weighted k-means clustering method. Chapter 6 is dedicated to our proposed weighted k-means clustering computational algorithm. In Chapter 7, we apply the newly improved weighted k-means method to both simulated datasets and real-application datasets with the results compared to those obtained with existing methods. Finally, directions for future work are laid out in Chapter 8.

Chapter 2

Cluster Analysis

Cluster Analysis or clustering is a generic label for a variety of procedures designed for unsupervised classification. Cluster Analysis identifies and classifies objects into different groups, so called “clusters”, based on the similarity or dissimilarity of a set of features the researcher concerned, or more precisely, partitions a data set into subsets, so that the data in each subset share some common features. The result of cluster analysis is a number of groups, which there are substantial differences between the groups, but strong similarities within a group. The early study of cluster analysis can be referred to R. C. Tryon (1939) who concerned individual difference in his psychology research. And later, from mid-1950’s, he used cluster analysis to classify social area and improved the theory. Hierarchical clustering (Ward 1963; Johnson 1967) and partitional clustering (Steinhaus 1957; MacQueen 1967) are the two major types of cluster analysis. Now cluster analysis is a very important and useful technique for exploratory data analysis, widely used in many fields, such as machine learning, data mining, pattern recognition, image analysis, document retrieval and bioinformatics.

2.1 Notations

To help our readers better understand the ensuing discussions, we have provided major notations on cluster analysis used throughout this thesis below.

m	Number of features or variables
N	Number of objects
k	Number of clusters
X_i	*ith* object with m features: $X_i = (x_{i1}, x_{i2}, \dots, x_{im}), \quad i = 1, 2, \dots, N$

$D(X_1, X_2)$ Distance between object X_1 and X_2

C_g j^{th} cluster center: $C_g = (c_{g1}, c_{g2}, \dots, c_{gm}), \quad g = 1,2, \dots, k$

D Distance matrix: $D = \{d_{ij}\}_{N \times N}, \ d_{ij} = D(X_i, X_j)$

W Variable Weighting vector: $W = (w_1, w_2, \dots, w_m)$

Z_i Feature-wise standardized i^{th} object

2.2 General Procedure of Cluster Analysis

A universal cluster analysis generally includes the following five steps (Jain and Dubes 1988):

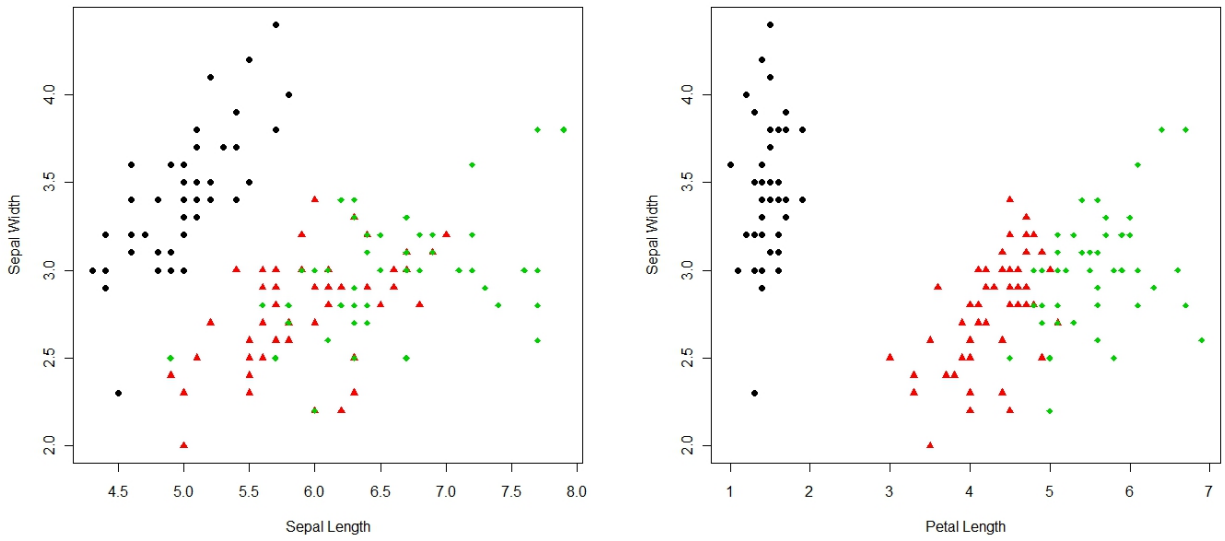
1) Pattern representation
2) Definition of similarity/dissimilarity measure appropriate to the data domain
3) Clustering process with specified algorithm
4) Data abstraction
5) Output validation

Cluster analysis is an exploratory tool and usually followed by other analytical techniques as the next step. Steps 4 and 5 described above mainly serve for future confirmatory analysis and are therefore not mandatory for cluster analysis. Usually, a clustering algorithm refers to the first three steps. Figure 2.1 below shows a typical work flow of the first three steps (Jain, et al. 1999).

Figure 2.1 General procedure in cluster analysis (Jain, et al. 1999)

Pattern Representation

The first step is pattern representation. It refers to the step to determine the number of clusters (for the partitional clustering analysis methods), the number and type of the features available and also relevant to the clustering problem. As an exploratory data analysis method, without knowing the “ground truth” behind, cluster number determination in cluster analysis could be very difficult in some cases. So far, there are no theoretical guidelines to suggest the appropriate cluster number in any specific situation. Indeed, this process is not fully controllable yet. The most common methods to determine the cluster number is through experts’ experience or simple descriptive statistics and graphic tools. However, such approaches are very subjective and, sometimes, may yield to a situation where true structure is hidden. Figure 2.2 provides a simple example. The figure is plotted from the well-known Iris data which we will discuss more in Chapter 7.



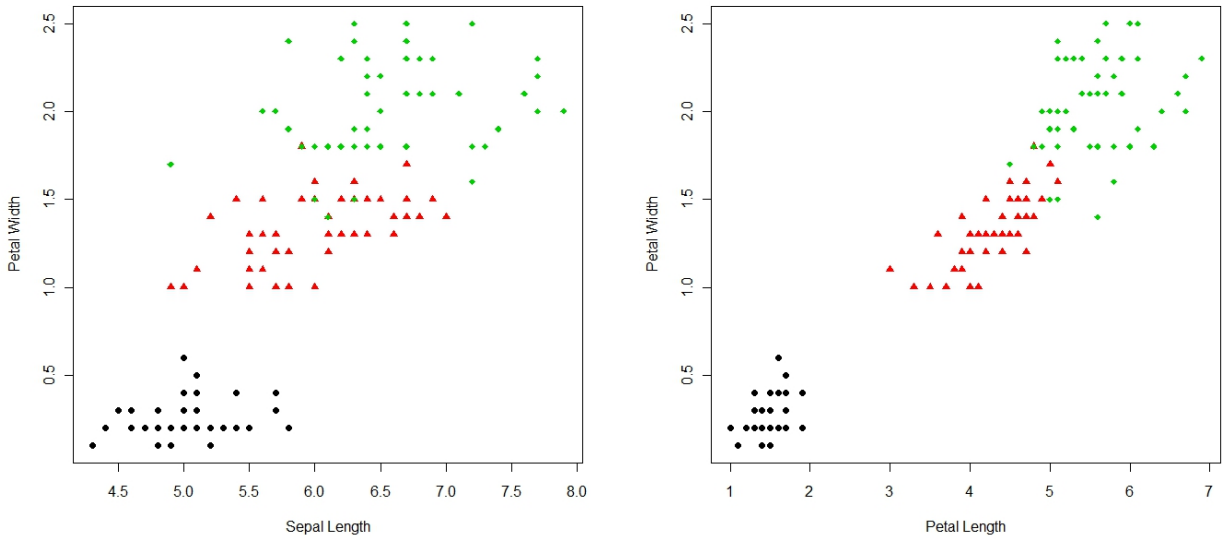


Figure 2.2 Iris four-dimensional data with three groups

In the Iris data, there are three classes representing three species of Iris. However from the graph, we may mistakenly consider that two clusters may be more appropriate. In hierarchical clustering, determination of cluster number does not need to be done at the beginning of the analysis which renders hierarchical clustering analysis relatively easier compared to the partitional clustering methods, which will be discussed later in this chapter.

Feature extraction and feature selection are used to find the most relevant features for the task. Feature extraction is to generate new features from the original feature set while feature selection is to identify a subset of features. Again, due to the exploratory nature of cluster analysis, feature extraction and feature selection become very difficult and usually done *ad hoc* as a part of the data abstraction.

Definition of Similarity/Dissimilarity

Similarity is to measure how similar two objects are in terms of falling into the same cluster. Cluster is defined based on the similarity measurement. The similarity measurement is essential to most clustering algorithms. In practice, it is more common to calculate the dissimilarity between two objects instead of similarity. Dissimilarity is the opposite of similarity. If two objects are close and highly likely to be clustered together,

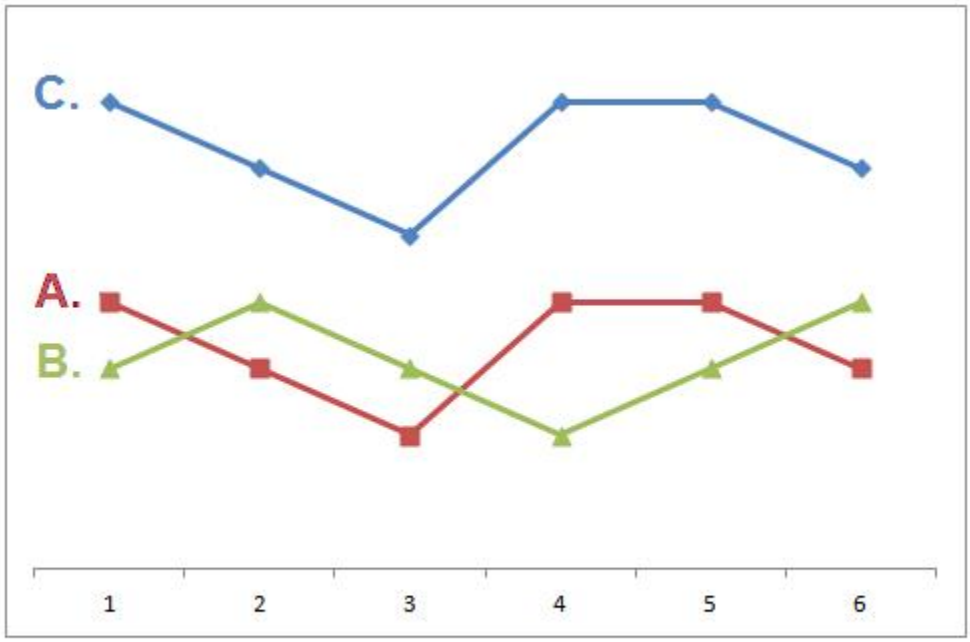
then the similarity between them should be large and the dissimilarity should be low, and vice versa. Dissimilarity is usually calculated using a distance function defined on the feature space. Because of the diversity of features, which could either be quantitative features, such as continuous values or discrete values, or qualitative features, such as nominal variables or ordinal variables, the distance function must be chosen carefully and appropriate for the specific feature space.

Table 2.1 Common distance function for continuous variables

Manhattan Distance	$D_{Man}(X_1, X_2) = \sum_{i=1}^{m} \lvert x_{1i} - x_{2i} \rvert$
Euclidean Distance	$D_{Euc}(X_1, X_2) = \left[\sum_{i=1}^{m} (x_{1i} - x_{2i})^2\right]^{\frac{1}{2}}$
Minkowski Distance of order *p*	$D_{Min}(X_1, X_2, p) = \left[\sum_{i=1}^{m} \lvert x_{1i} - x_{2i} \rvert^p\right]^{\frac{1}{p}}$
Chebyshev Distance	$D_{Che}(X_1, X_2) = \lim_{p \to \infty} D_{Min} = \max_i(\lvert x_{1i} - x_{2i} \rvert)$
Mahalanobis Distance	$D_{Mah}(X_1, X_2) = [(X_1 - X_2)^T \Sigma^{-1} (X_1 - X_2)]^{\frac{1}{2}}, \Sigma = cov(X_1, X_2)$
Correlation-based Distance	$D_{Cor}(X_1, X_2) = \frac{1 - corr(X_1, X_2)}{2}$

A lot of distance metrics (Huttenlocher, et al. 1993; Xing, et al. 2003) has been used in cluster analysis as similarity measurement. The most commonly used distance metric for continuous features is the Euclidean distance (Per-Erik 1980) which is a special case of the Minkowski metric. Mahalanobis distance is used if the correlation between the features may distort the distance measure. A list of distance functions for continuous variable has been shown in Table 2.1. Practitioners also developed distance metrics for discrete features or qualitative features (Ichino and Yaguchi 1994; Wilson and Martinez 1997; Finch 2005). Different distance functions may yield different results for the same pair of objects. Subsequently, same clustering algorithm with different distance measure could yield dramatically different clusters. For example, in Figure 2.3,

we show three objects A, B and C with 5 continuous features. Euclidean distance function concludes the distance between A and B is smaller than the distance between A and C, but correlation distance function gives the opposite conclusion. In cluster analysis, it is very critical to know the appropriate distance function and understand the difference.



$$D_{Euc}(A,B) = 3 < 7.3 = D_{Euc}(A,C)$$
$$D_{Cor}(A,B) = 0.7 > 0 = D_{Cor}(A,C)$$

Figure 2.3 Comparison between Euclidean distance and correlation-based distance

Clustering Algorithm

After defining the appropriate distance, or similarity measurement, the next step is to choose the right clustering algorithm. Given the popularity and usefulness of cluster analysis, hundreds of clustering algorithms have been proposed in the literature (Jain, et al. 1999; Jain, et al. 2000; Steinley 2006). However, as previous mentioned, most of algorithms belong to two major classes: hierarchical clustering, and partitional clustering.

In hierarchical clustering, a nested series of partitions are produced. When performing a hierarchical clustering, users do not need to determine the cluster number

at the beginning. Instead, the best partition plus the corresponding cluster number can be selected from the series of partitions afterwards. In hierarchical clustering, one can also choose either agglomerative (bottom up) approach or divisive (top down) approach combined with different ways of measuring cluster centers/distances such as the single linkage algorithm, complete linkage algorithm or average linkage algorithm (Ward 1963; King 1967). Detailed discussion on hierarchical clustering will be discussed later in this chapter.

In partitional clustering, only a single partition of the data is obtained instead of a group of partitions. As a consequence, the cluster number must be decided before performing the partitional clustering. As we discussed in Chapter 1, cluster number is always difficult to decide. Milligan and Cooper (Milligan and Cooper 1985) reviewed some ideas to help making this choice. Partitional clustering algorithms usually produce the clusters by optimizing a defined criterion function. A partitional algorithm can be classified as a hard algorithm if each object can only be allocated to a single cluster, such as the K-means algorithm (MacQueen 1967); or a soft algorithm when each object can be assigned to several clusters with degrees/probabilities of membership, such as the fuzzy c-means clustering (FCM) (Ruspini 1969; Bezdek, et al. 1984; Bezdek, et al. 2005).

While Hierarchical clustering has only agglomerative algorithm and divisive algorithm, partitional clustering has various choices in algorithm (Jain, et al. 1999). Without any distribution assumptions, one can apply nonparametric approach, density-based clustering (Jain and Dubes 1988), such as Nearest Neighbor clustering (Yianilos 1993; Kenward, et al. 2001), DBSCAN (Ester, et al. 1996; Arlia and Coppola 2001) and OPTICS (Ankerst, et al. 1999). Then, with mixture Gaussian assumption, such as Expectation Maximization (EM) algorithm (Dempster, et al. 1977; Redner and Walker 1984; Peel and McLachlan 2000), and Cross Entropy (CE) algorithm (Botev and Kroese 2004; Rubinstein and Kroese 2004) also have been applied to solve partitional clustering problems. Most recently, subspace clustering and correlation clustering have been developed specifically for high-dimensional data to deal with the curse of dimensionality (Agrawal, et al. 2005; Kriegel, et al. 2009). In general, K-means

clustering is the simplest and most commonly used partitional clustering algorithm thanks to its considerable efficiency, which will be further discussed later in this chapter.

Data Abstraction and Cluster Validation

If the goal of a study is just to discover the number of clusters or the structure in a data set, then a partition of the data set is the end product and there is no need to do data abstraction or cluster validation. However, in most real world applications, cluster analysis is usually the first step to explore the data, and then other statistical methods or data analysis techniques will be applied either on each cluster separately, or on the cluster centers/seeds. In cluster analysis, data abstraction is used to extract a simple and common representation of each cluster. The most popular way is to use of the cluster center to represent each cluster (Diday and Simon 1976). The first principal component has also been used as representation of a cluster in many fields (Zhang, et al. 2008). The best way of cluster representation depends on the application itself as discussed in Duran and Odell (1974).

Cluster validation is used to evaluate the performance and assess the output of a clustering algorithm. There are two types of validation. In external evaluation, the clustering output is assessed using external data which was not used for clustering (e.g. class labels if available; external benchmarks). Some external criterions include Rand measure (Rand 1971), F-measure (Manning, et al. 2008), Jaccard index (Hamers, et al. 1989) and Confusion matrix (Townsend 1971). External evaluation methods evaluate the clustering output with extra knowledge. However, the recovery of known knowledge may not necessarily to be the primary intention of exploratory data analysis. In internal evaluation, by contrast, the output is evaluated, based on the data used for clustering, by determining if the output is essentially appropriate for the data. However, as pointing out by Manning (2008), a high score in internal evaluation do not necessarily result in effective information recovery. Commonly used internal criterions include Davies-Bouldin index (David and Donald 1979) and Dunn index (Dunn 1974). For interests in this direction, please see the detailed discussion in (Jain and Dubes 1988)

2.3 Hierarchical Clustering

Hierarchical clustering algorithm is also called connectivity based clustering, which will produce a hierarchy of partitions. The output of a hierarchical clustering is a tree structure called dendrogram representing the nested grouping of objects and similarity levels. An example of dendrogram is shown in Figure 2.4. By cutting the dendrogram at different levels (Height in Figure 2.4), different clusters of the data can be obtained.

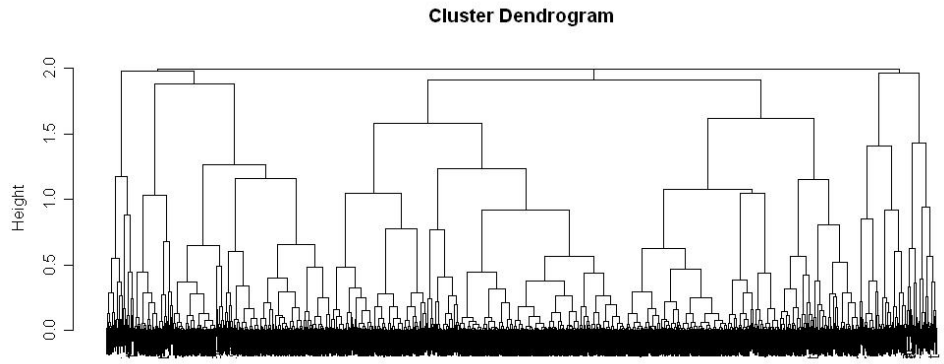


Figure 2.4 Dendrogram

As shown in Figure 2.5, the first step in hierarchical clustering is to define the similarity/dissimilarity using distance measure as we mentioned before. Hierarchical clustering is very flexible in selecting distance functions. However, different distance functions identify different features in the data and thus the dendrogram structure using different distance functions may naturally differ.

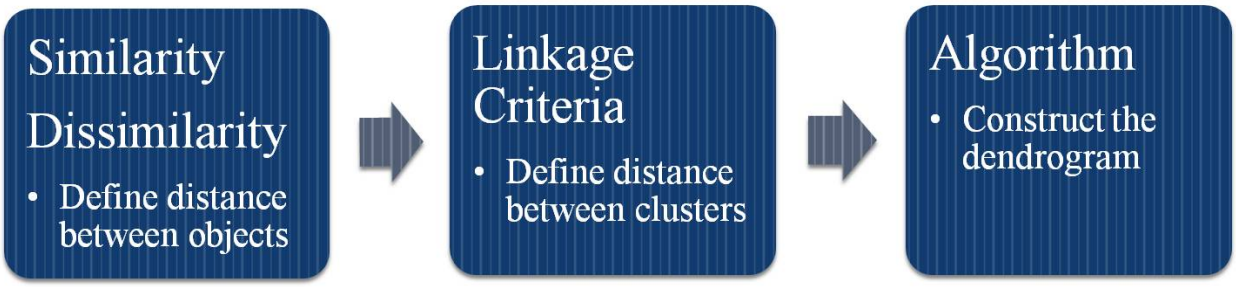


Figure 2.5 General procedure of hierarchical clustering

When a distance function is determined, the distance is properly defined between any two objects. Then hierarchical clustering offers several linkage criteria to define the distance between two sets of objects. Some commonly used linkage criteria are listed below:

Single linkage: also called minimum linkage, defined as the minimum distance between any two objects from two sets.

Complete linkage: also called maximum linkage, defined as the maximum distance between any two objects from two sets.

Average linkage: also called mean linkage or UPGMA, defined as the mean of the distance of all paired objects from two sets.

Ward's criterion: in ward's criterion, the distance between two sets of objects is defined as the increase in variance if two sets of objects are merged.

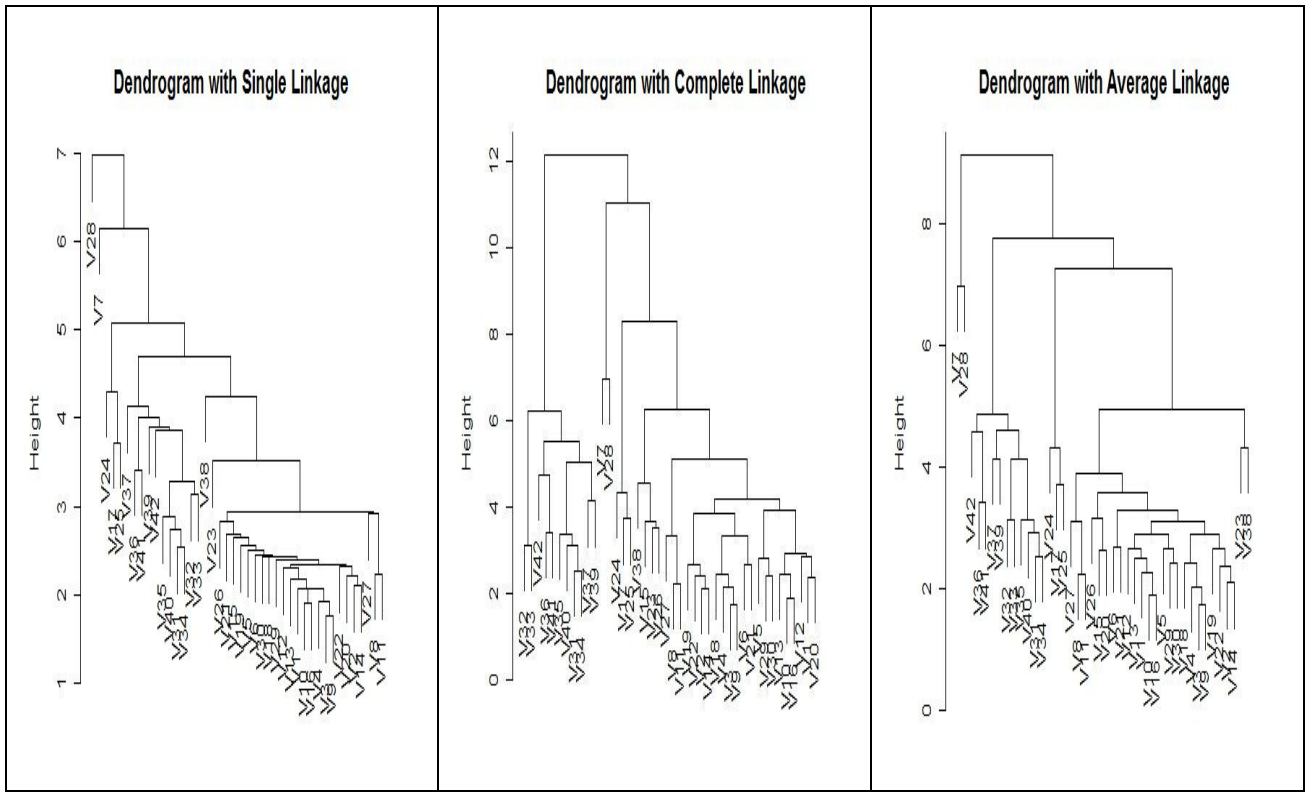


Figure 2.6 Dendrograms with three difference linkage criteria using same distance function on same data

For the same data, different linkage criteria will lead to different tree structure, even with the same distance measure (Figure 2.6). Ward's criterion only works well if the data is approximately normal. It has been reported that single linkage always results in a chaining effect (Nagy 1968) as well as average linkage frequently produces the snowballing effect (Huth, et al. 1993; Serrano, et al. 1999). Although such limitations can be partially eliminated by stopping the clustering process at different level of dissimilarity for different parts of data, complete linkage is still preferred and the most commonly used in many applications. It has been observed that complete linkage hierarchical clustering produces tight and compact clusters as well as more meaningful dendrogram than the single linkage method (Jain and Dubes 1988; Baeza-Yates 1992).

Hierarchical clustering is a "step by step" algorithm allows either building up (agglomerative), or breaking down (divisive), the hierarchy of clusters (Figure 2.7). The agglomerative algorithm, also called "bottom-up" approach, begins at the bottom of the tree with each object as a single cluster. Then in each step, the closest two clusters, measured by the distance metrics and selected linkage, are merged as a larger cluster until all elements are in one cluster. The divisive algorithm, also called "top-down" approach, on the other hand, has the reverse order. It begins at the top of the tree with a single cluster including all the objects and split is produced in each step recursively until no further splits any more. In practice, agglomerative algorithm is more popular compared to divisive algorithm because of the complexity. The complexity of divisive algorithm is $\boldsymbol{O(2^n)}$ while the complexity of agglomerative algorithm is $\boldsymbol{O(n^3)}$ in general and $\boldsymbol{O(n^2)}$ for some special cases.

Hierarchical clustering provides a wide choice on cluster number, there is no needs to determine the cluster number in advance. Each level in the dendrogram provides a unique partition of the data and the final clusters can be decided by comparing all the possible results. More details about determining cluster number will be introduced later. Dendrogram provides very high interpretability of the whole procedure, which makes hierarchical clustering a very popular choice. However, on the other hand, such tree structure is very sensitive and instable. Different linkage methods, a small change in the data, or a perturbation at early steps, can results in significant

difference in dendrogram. More important, the hierarchy structure imposed by hierarchical clustering may not actually exist in the data.

Figure 2.7 Divisive and agglomerative hierarchical clustering

2.4 K-means Clustering

K-means clustering is the earliest and most commonly used partitional clustering algorithm in the literature “from more than a dozen different fields” (Steinley 2006). Back to 1960’s, many researchers (Thorndike 1953; Cox 1957; Fisher 1958; Engelman and Hartigan 1969) suggested to partition the data by minimizing within-group variation so that the final partitions produced can reflect a certain level of homogeneity within clusters and heterogeneity between clusters. Then the terminology “K-means” was first used by James MacQueen in 1967. However, the original idea was proposed by Hugo

Steinhaus (1957) in 1957. The basic idea is to assign all subjects into k clusters in which each subject belongs to the cluster with the nearest center represented as the mean of all the objects in the cluster. Unlike hierarchical clustering, k-means clustering requires determining the cluster number k in advance and only produces a single partition with k clusters. When cluster number k is fixed, K-means clustering is actually an optimization problem to find the best k subgroups with k cluster centers C_g by minimizing the sum of within group sum of squares (WGSS) as follow:

$$\sum_{g=1}^{k}\sum_{i\in I_g}\sum_{j=1}^{m}(x_{ij}-c_{gj})^2 \tag{2.1}$$

The standard K-means algorithm, also known as Lloyd's algorithm (Figure 2.8), was originally proposed by Stuart Lloyd in 1957 and first published in (1982). In Lloyd's algorithm, first, k cluster centers are initialized randomly and each subject is assigned to the nearest center. Then in each step, k cluster centers are re-calculated based on the assignment and all subjects are re-assigned to new clusters until the k cluster centers remain no change.

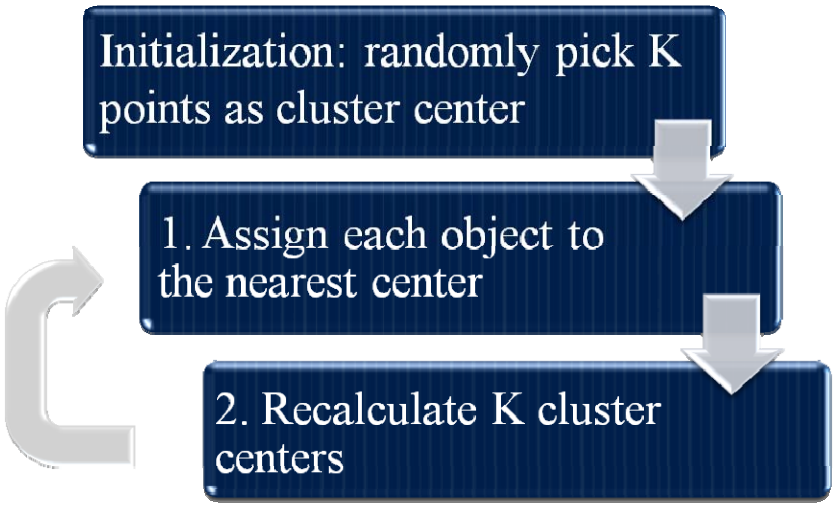


Figure 2.8 Lloyd's algorithm

With additional assumption that each cluster follows a multivariate normal distribution, k-means problem can be solved by estimating a finite Gaussian mixture model. Then either Expectation-Maximization algorithm (Dempster, et al. 1977; Redner

and Walker 1984; Peel and McLachlan 2000) or Cross-Entropy algorithm (Botev and Kroese 2004; Rubinstein and Kroese 2004) is used to fit the model and each object is assigned to the cluster with highest probability. These distribution-based algorithms outperformed the standard algorithm in specific data structures. However, it has been shown that EM algorithm has the issue to identify the within-in cluster covariance matrices when solving a k-means problem (De Backer and Scheunders 1999) and in high dimensional data, even dimension>2, both EM and CE algorithms will possibly generate spurious clusters (sometimes called degenerate clusters) (McLachlan and Peel 2000) Furthermore, several alternatives have been reported in the literature using K-median (Kaufman and Rousseeuw 1990; Park and Jun 2009), K-midranges (Späth 1985; Carroll and Chaturvedi 1998) or K-modes (Huang 1998; Chaturvedi, et al. 2001; Huang and Ng 2003), instead of K-means, as k cluster centers. However, these k-means-like algorithms only works well as expected in special cases (Carroll and Chaturvedi 1998; Garcia-Escudero and Gordaliza 1999), but they shared the same limitations presented in original K-means clustering (Huber 1981; Arabie and Hubert 1994; Chaturvedi, et al. 2001).

A well-known problem of k-means clustering is that it may fail to provide the global optimum and very sensitive to the randomly initialized centers (Steinley 2003). Even through, it has been reported that k-means algorithm usually shows very good cluster recovery properties (Dimitriadou, et al. 2002; Steinley 2003). K-means clustering is favorable because it is easy to implement and very fast with time complexity of ***O(n)***. Compared to hierarchical clustering, it makes k-means clustering unattractive to determine cluster number k in advance. However, given the low time complexity, it is possible to run K-means algorithm with a range of cluster number k and selected the most appropriate one afterward, just like the way hierarchical clustering does.

2.5 Application of Cluster Analysis

Cluster analysis, as the major unsupervised learning technique, has been heavily used in many disciplines to discover the ground truth, recover hidden information and explore the unknown portion of the world.

In biology, cluster analysis has been applied in transcriptomics, evolutionary biology and bioinformatics. In transcriptomics, cluster analysis is used to build groups of genes with gene expression patterns (Subramanian, et al. 2005). These groups often contain functionally related proteins. Varela, et al. (2011) performed hierarchical clustering for transcriptome, proteome and endometabolome to study the biology of win yeast in 2011. In evolutionary biology and bioinformatics, cluster analysis is widely involved into the studies related with high-throughput genotyping platforms (Hormozdiari, et al. 2009), microarray data analysis (Eisen, et al. 1998) and protein-protein interaction network (Ito, et al. 2000; Uetz, et al. 2000; Chua, et al. 2006; ,55-59). Andreopoulos and colleagues (2009) reviewed around 40 clustering algorithms applied in bioinformatics. A separate section will be provided in next chapter to discuss the application of cluster analysis in microarray data analysis.

In ecology, cluster analysis is used to reveal biogeographical or temporal patterns by clustered patterns of molecular sequences (Whitaker, et al. 2003). Another application is to find out clones or variants from environmental samples based on genetic features or phenotypic markers (Acinas, et al. 2004). Ramette Alban (2007) conducted a review on statistical analysis in microbial ecology. The review includes 7748 publications which are published between 1900 and 2006. Around 40%-50% publications used the word “cluster analysis” in the titles or abstracts.

In medicine research, cluster analysis is mainly used in medical imaging study to analyze the data obtained from Functional MRI (Goutte, et al. 1999; Baumgartner, et al. 2000; Cordes, et al. 2002) and PET (Wong, et al. 2002; Guo, et al. 2003; Kamasak and Bayraktar 2007). O’Sullivan (1993) used partitional clustering to transform a set of tissue time-activity curve (TAC) into groups of homogeneous TAC. Liptrot and collaborators (2004) applied a two-stage hierarchical k-means clustering algorithm on

PET time series to extract a cerebral vasculature ROI and built a kinetic model of the brain. Also clustering analysis has been used to improve the Signal-to-Noise Ratio (SNR) of dynamic PET data (Kimura, et al. 1999). Recently, Robinette's team (2009) applied hierarchical clustering and biclustering algorithm on nuclear magnetic resonance (NMR) imaging data to profile the changes in biofluid metabolic composition. Desai and colleagues (2011) used cluster analysis to analyze cytokine profile in severe asthma subphenotypes.

In business and marketing, cluster analysis has been used starting from 1960s (Punj and Stewart 1983; Arabie and Hubert 1994). The primary application of cluster analysis in marketing is for market segmentation (Wind 1978; Saunders 1980; Spiller and Lohse 1997). Another important use of cluster analysis in business and marketing is to understand clients' behaviors by grouping homogeneous clients (Kiel and Layton 1981; Desarbo, et al. 1991) and subsequently make marketing decisions and strategies (Flavián and Polo 1999). Most recent, Hosseini's group (2010) developed a customer relationship management (CRM) method using k-means clustering to assess customer; Zhang and other researchers (2011) employed Kohonen clustering (Tsao, et al. 1994) algorithm to study telecom customers' behavior.

In computer science, cluster analysis is an essential tool to handle large data. Jain (1999) explained the importance of clustering in image segmentation (Frigui and Krishnapuram 1999; Tung, et al. 2010), information retrieval (Carpineto and Romano 1996; Bhatia and Deogun 1998; Messai, et al. 2008) and data mining (Judd, et al. 1998). For example, in image segmentation, cluster analysis is used to group an input image into homogeneous regions based on some image-specific features. Cluster analysis is also widely used in pattern recognition (Ester, et al. 1996; Hinneburg and Keim 1998; Jain, et al. 2000; Han, et al. 2006) and image processing (Bagui 2005; Bezdek, et al. 2005; Gerlinger, et al. 2009)

In atmospheric sciences, cluster analysis is applied to detect circulation regimes and classify weather patterns (Mo and Ghil 1988; Michelangeli, et al. 1995; Yiou and Nogaj 2004). Both hierarchical clustering (Cheng and Wallace 1993; Mote 1998; Vrac, et al. 2007) and k-means clustering (Brinkmann 1999; Solman and Menéndez

2003; Santos, et al. 2005; Esteban, et al. 2006) are used in this kind of study. Kalkstein's team (1987) performed a comparison of different clustering techniques in clustering weather types. Gong and Richman (1995) discussed the performance of cluster analysis in climate regionalization and the effect of different distance measurements.

Furthermore, cluster analysis has also been employed in other fields as well. Clatworthy and co-workers (2005) reported a thoughtful review on the application of cluster analysis in health psychology in last decade. Basak with colleagues (1988) used cluster analysis to find the structural similarity of 3000 chemical compounds. Zhang and Maringer (2010) combined clustering technique with traditional asset allocation methods in modern portfolio management. With this new approach, they improved portfolio stability and resulted in higher risk-adjusted returns.

2.6 Recent Development

With the wide applications of cluster analysis, new algorithms are proposed constantly and considerable effort has been put on improving the performance of hierarchical clustering and k-means clustering. Zhang, Ramakrishnan and Livny (1996) proposed BIRCH which shorted the runtime and improved the efficiency compared to other hierarchical algorithms. This algorithm received the SIGMOD 10-years test of time award due to the improvement. Cheng and Church (2000) created "biclustering" which performs hierarchical clustering on both object and feature level simultaneously. Kanungo's group (2002) designed a filtering algorithm for k-means clustering with increased efficiency as the separation between clusters increases. More recently, fuzzy or overlapping clustering drawn many attentions by allowing each object belong to multiple clusters while clusters are mutually exclusive in classical cluster analysis (Banerjee, et al. 2005; Eyke 2005). Knowledge-based clustering, which incorporating extra background knowledge into clustering, becomes an interesting topic in bioinformatics while classical cluster analysis is purely data-driven (Hanisch, et al. 2002; Pan 2005; Tseng 2007). Furthermore, several papers discussed the variable

weighting techniques to improve the performance and retrieved more information from the data (Modha and Spangler 2003; Tseng 2007; Shen, et al. 2010).

Besides numerical data clustering, some recent developments are related to categorical data. Huang (1998) extended the k-means algorithm to categorical data. Kim and colleagues (2004) developed fuzzy clustering, while Guha's team (2000) and He's group (2002) developed computer-based algorithms for categorical data. Andritsos with other researchers (2004) created a new algorithm named "LIMBO" with a novel distance measure for categorical data and improved scalability of other hierarchical clustering algorithms. Cluster analysis for numerical and categorical mixed data has also been developed in the recent years (Chiu, et al. 2001; Li and Biswas 2002; Hsu, et al. 2007).

In this thesis, we shall discuss two new developments on cluster analysis. In the first part, we shall extend the modern multiple-objective cluster analysis from using a single set of features to multiple distinct sets of features by developing the novel compound clustering method and the constrained clustering method. In the second part, we theoretically derive the global-optimal guaranteed variable weights based on the method of Lagrange multiplier and the Karush-Kuhn-Tucker conditions. Then we shall propose a novel algorithm for the weighted k-means clustering to improve the current weighted k-means clustering method (Huh and Lim 2009). Numerical examples on both simulated and real data are provided at the end to illustrate our method.

Part I. Modern Multi-objective Cluster Analysis: Compound Clustering and Constrained Clustering

Chapter 3

Existing Multi-objective Cluster Analysis

3.1 Cluster Analysis in Microarray Data

High throughput techniques are very important nowadays in many areas of biology research. Microarray technique is one of such high throughput techniques to generate large-scale gene expression data and enable the biological investigation to conduct in the gene level, where structural information about protein sequence and regulatory information about protein expression are stored. Microarray data is very useful to biological research for identifying differentially expressed genes, function annotation of coexpressed genes, regulatory mechanisms and diagnostic. The high dimensional, large scale nature of microarray data increases the demand of advanced and sound statistical methodology.

Many statistical methods have been applied on microarray data to identify differentially expressed genes in high dimensions (Kerr, et al. 2000; Nadler, et al. 2000; Lin, et al. 2003). For study focused on gene co-expression, functional annotation and coregulation, cluster analysis is widely used as the first-step statistical analysis to group genes into sets with similar expression patterns. In the last decade, clustering algorithms in microarray analysis have been extensively reviewed (Dysvik and Jonassen 2001; Boutros and Okey 2005; Gollub and Sherlock 2006), compared and validated (Datta 2001; Kerr and Churchill 2001; McShane, et al. 2002; Gat-Viks, et al. 2003). Generally, if the researcher has a desired cluster number by external knowledge, partition clustering, such as K-means clustering or fuzzy C-means clustering (Dembélé and Kastner 2003) are preferred due to the time efficiency. If cluster number is unknown, hierarchical clustering is better because of the flexibility on cluster number. People

usually use normalized Euclidean distance or correlation based distance as well as partial correlation based distance (Waddell and Kishino 2000) for hierarchical clustering.

However, those popular algorithms have limitations when applying to microarray data. From statistical point of view, hierarchical clustering is embedded with the assumption that the internal structure of the data is essentially hierarchical, which implies a hierarchy of genes. However, if this is not the case, or genes are not correlated across all levels, the performance of hierarchical clustering may be inferred. Also in partitional clustering, the estimated cluster number may not represent the inherent one in the data. In that case, such bias will be taken into the clustering procedure and the result may be inferred as well. Lots of research have been conducted in this area to improve the performance of clustering algorithms for microarray data (Yeung, et al. 2001; Hanisch, et al. 2002; Ernst, et al. 2005; Chipman and Tibshirani 2006).

More important, from biology point of view, the biological fundamental behind cluster analysis is that co-expressed genes are always co-regulated, that is, genes with similar expression are supposed to be involved in similar biological processes. However, it has been reported in the literature that the similarity in gene expressions may not necessarily reveal biological similarity (Clare and King 2002; Gibbons and Roth 2002), also genes involved in the same biological process are not always perfectly correlated (DeRisi, et al. 1997). To overcome this difficulty and find gene clusters better presenting biological process, several novel clustering algorithms has been proposed. Existing literature includes Multi-objective clustering and knowledge-based clustering.

3.2 Multi-objective Cluster Analysis

Traditional cluster analysis algorithms always only use a single criterion or objective -- such as the objective of K-means clustering is the compactness of objects and the objective of hierarchical clustering is the connectedness of the objects. This renders clustering of data with several major features difficult (Figure 3.1). To overcome

this un-necessary limitation, some recent works proposed the new idea of multi-objective clustering (Handl and Knowles 2004, 2005; Cheng, et al. 2006; Korkmaz, et al. 2006; Mitra and Banka 2006; Di Nuovo, et al. 2007; Handl and Knowles 2007) applying multiple cluster criteria (objective functions) simultaneously to one signle data set. An alternative approach is clustering ensembles (Strehl and Ghosh 2003) featuring a posteriori combination of different clustering results by means of ensemble methods. The multi-objective approaches are more robust and stable than the traditional single objective cluster methods. Similar to hierarchical clustering, they produce a set of partitions as the final result, and from this set, users can choose the result most suitable for their particular application (Verma and Blumenstein 2008). Most cluster quality validation procedures also determine the cluster number at the same time. A notable algorithm is the automatic k-determination scheme (Handl and Knowles 2007; Matake, et al. 2007) inspired by the Gap statistic (Tibshirani, et al. 2001).

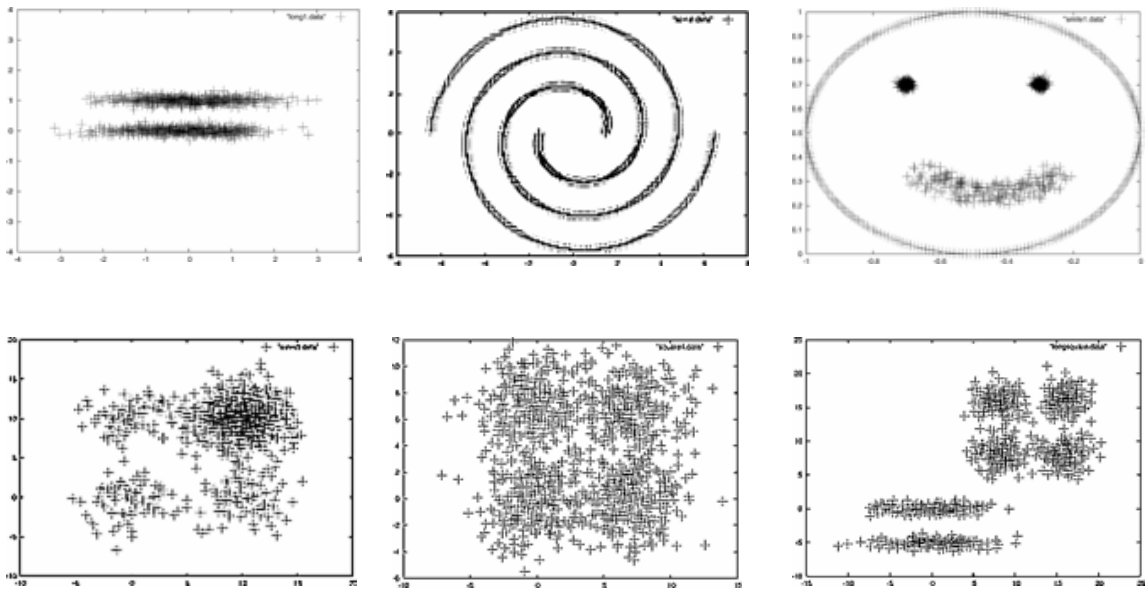


Figure 3.1 Motivation of modern multiple-objective clustering analysis: synthetic 2D data sets exhibiting a wide range of different data properties (Handl and Knowles 2007).

Most multi-objective clustering approaches use the Multi-Objective Evolutionary Algorithm (MOEA) obtained via the framework of Pareto optimality, or Pareto efficiency. Named after Vilfredo Pareto, an Italian economist who used the concept in his studies

of economic efficiency and income distribution, the Pareto optimality has broad applications in economics, game theory, engineering and the social sciences. We now illustrate this optimization method using terms from its economic roots. Given a set of alternative allocations, a change from one allocation to another that can make at least one individual better off without making any other individual worse off is called a Pareto improvement. An allocation is Pareto efficient or Pareto optimal when no further Pareto improvements can be made.

In the context of multiple-objective clustering, after defining the objectives of data clustering, MOEA usually has two steps: initialization and optimization. Initialization will build an initial partition of the data. The minimum spanning tree (MST) (Handl and Knowles 2007; Matake, et al. 2007) is perhaps the most widely used method for initialization. After the initialization, the Pareto optimization algorithm will be performed to minimize the objective functions. Theoretically, multiple objective functions can be applied, but thus far, only two objectives have been examined in papers published to date: one is the connectivity and the other is the overall deviation. These objective functions are defined similarly across papers with slight changes. Connectivity evaluates whether the most similar objects have been placed in the same cluster. The overall deviation expresses the within-cluster compactness by evaluating the overall summed distances between objects and their corresponding cluster center. Connectivity is minimized by decreasing the cluster number while the overall deviation is minimized by increasing the cluster number. Thus the optimization is achieved by balancing these two objective functions. Many multi-objective optimization algorithms can be applied, such as MOGA (Deb 2003), VIENNA (Handl and Knowles 2004), MOCK (Handl and Knowles 2007; Matake, et al. 2007) which is an improvement from PESA-II (Corne, et al. 2001). These algorithms utilize different combinations of operators such as genetic representation, uniform crossover and neighborhood- biased mutation, to optimize the initial partition.

In microarray data analysis, multi-objective approach has recently been used to explore the gene clusters which are not only similar in expression profiles but also connected in biological process. Fleury (2002) applied multi-objective optimization to

reveal temporal patterns in mouse retinal genes. Divina and Aguilar-Ruiz (2007) extended the popular biclustering techniques with multi-objective approach. Bandyopadhyay (2007) proposed a two-stage clustering algorithm with Fuzzy C-means and Multi-Objective genetic algorithms. Faceli and colleagues (2009) combined cluster ensembles and multi-objective clustering and applied for gene expression data analysis to handle different types of clusters existing in gene expression data.

In next chapter, we will propose two new multi-objective clustering methods, independent from above MOEA, and apply on a microarray study collaborated with Cold Spring Harbor laboratory with two objectives: 1) maximize the similarity of gene expression; 2) maximize the correlation of biological functions.

3.3 Knowledge-based Cluster Analysis

Cluster analysis is a data-driven unsupervised learning method. It tries to explore the underlying structure from the data completely, without any pre-knowledge or information. However, as we discussed before, gene clusters obtained from traditional cluster analysis on expression data may not fully uncover the biological meaning behind it. Multi-objective clustering approaches still didn't solve this problem by focusing on gene expression data alone. Given the availability of the various sources of biological data, it becomes popular to incorporate biological knowledge into cluster analysis to improve the performance and retrieve biological meaningful clusters (Cheng, et al. 2004; Huang and Pan 2006). Such approaches are called knowledge-based clustering.

Gene Ontology (GO) (Ashburner, et al. 2000) is usually used as the prior biological knowledge in microarray data analysis (Dahlquist, et al. 2002; Zeeberg, et al. 2003; Pan 2005). GO is a hierarchical classification structure displayed as a direct acyclic graph (DAG) (Wong, et al. 2002). In DAG, each GO node representing a biological process and the end notes are usually the genes. Usually a child node is a part of the parent node or a specific case. If two genes share more patient nodes, they

are more biological similar. GO annotation has been used in both hierarchical clustering and partitional clustering.

In hierarchical clustering, biological knowledge usually used to define the biological similarity between genes and then combined with gene expression similarity as overall distance metrics. The first biological similarity measure was proposed by Hanisch and collaborators in 2002. They defined the biological similarity using metabolic network on enzymes, KEGG database (Kanehisa, et al. 2011) and the overall distance as a sum of two logistic functions of both biological similarity and expression similarity. This method was criticized because the genes and enzymes are not one-on-one corresponding, which makes the result very difficult to interpret. From then, several algorithms were proposed using GO annotation as biological similarity. Cheng and other researchers (2004) proposed a biological similarity between two genes based on the common edges they shared in GO and proposed hierarchical clustering on biological similarity alone as well as the average of biological similarity and Euclidean distance on expression data.

In partitional clustering, biological knowledge is incorporated as a weighting parameter on expression-based distance measure. Huang and Pan (2006) proposed a two step K-medoids methods (Kaufman and Rousseeuw 1990), a robust version of k-means clustering with a scaled expression distance metric with the nonnegative scale parameter $\gamma < 1$ if two genes have a common biological function indicated by GO or $\gamma = 1$, otherwise. Tseng (2007) proposed a PW-K-means algorithm with a penalty term and a cluster weighting factor defined as a logistic function of prior GO biological information, which is similar with Hanisch's 2002 model. Tari and colleagues (2009) extended fuzzy c-means clustering with variable weight obtained from GO annotation. Shen's team (2010) extended k-means clustering with a variable weight from GO information and allows for a set of scattered genes remaining un-clustered.

A common drawback on all existing methods is that only genes from GO or other biological database, such as KEGG, are used. Genes, which are not included in GO or other biological database, are assumed to have no function at all. However, this assumption is questionable. All the databases are knowledge-based, summarized from

previous experiments and research. A main goal of microarray analysis is to discover unknown gene functions. Such assumption will bias the results and limit our findings. Furthermore, in partitional clustering, by adding a weighting on cluster level, it may not result in the improvement on the gene-level similarity.

3.4 Cluster Number Determination

No matter which cluster method – hierarchical or partitional, single or multiple objective(s), finding the most reasonable cluster number is always critical. There is, unfortunately, no standard approach available (Girman 1994) to solve this problem. Cluster number determination is still a difficult and open question to date. The determination always involves, directly or indirectly, user experience and preference, and also the data property and visualization.

Lots of approaches have been proposed to find the most appropriate cluster numbers. Milligan and Cooper (1985) conducted an extensive review of 30 approaches used to determine the cluster number for hierarchical clustering. Recently, Salvador and Chan in 2004 summarized existing cluster number determination approaches into five categories. They are:

1) Cross-validation;
2) Penalized likelihood estimation;
3) Permutation tests;
4) Resampling;
5) Finding the knee of an error curve

They also pointed out that “The majority of these methods to determine the number of clusters/segments may not work very well in practice”. First four categories are all very computational intensive. These three methods require re-running the clustering many times, which is difficult for large dataset.

Dudoit and Fridlyand (2002) gave a review on majority methods to find the knee of an error curve:

1) Calinski and Harabasz (1974) index;

2) Krzanowski and Lai (1985) index;

3) Hartigans (1975) statistic;

4) Gap and gapPC (Tibshirani, et al. 2001).

For a given partition, these methods are aimed to locate the "knee" of an error curve, which is the point of maximum curvature as the appropriate number of clusters. However, Douglas Steinly (2006) criticized these methods only perform well if all clusters have similar size and shape.

Some other methods have been proposed as well. Sarle (1996) gave such a message: "If your purpose in clustering is dissection, that is, to summarize the data without trying to uncover real clusters, it may suffice to look at R^2 for each variable and pooled over all variables. Plots of R^2 against the number of clusters are useful." Kaufman and Rousseeuw (1990) proposed a method to determine the cluster number for partitional clustering using the largest average silhouette. Chiang and Mirkin (2006, 2007) compared eight most popular approaches to determining the cluster number for K-means clustering and concluded that silhouette produced the most consistent results.

In this thesis, we will develop a new statistics called "complete linkage" R^2 along with largest average silhouette to determine the cluster number determination for the compound and constrained cluster analysis, which will be introduced in the next chapter.

Chapter 4

Proposed Multiple-objective Clustering Approaches and Application in Microarray Data Analysis

Inspired by multi-objective cluster analysis which explores multiple features from a single data set and knowledge-based clustering which utilizes the multiple data sources available, in this chapter; we developed two novel multiple-objective cluster analysis method, the compound cluster analysis and the constrained cluster analysis, to cluster data using multiple data sources and multiple similarity measures. We illustrate our framework through a dual-objective and dual-data sources problem although the ideas are easily generalized to the multiple objective and multiple data sets scenarios. A real application in microarray data analysis is provided at the end.

4.1 Compound and Constrained Clustering

Cluster analysis is a data-driven technique. Traditional clustering algorithms consist of a single objective and a single distance metrics generated from a single data set. However, in reality, we may obtain heterogeneous data types from multiple sources which describe the same object from different views. Integrating all the information on hand could lead us to a better understanding of the object of interest. A very good example is the knowledge-based clustering we introduced in section 3.3. In biology, we have continuous microarray data to describe gene expression profile and also biological

network databases to describe the relationship between genes in terms of biological processes or functions. By integrating all the information, knowledge-based clustering achieves superior performance in grouping genes.

Here we developed two novel multiple-objective cluster analysis methods, the compound clustering and the constrained clustering, to cluster data by integrating multiple data sources and multiple similarity measures. We give the general compound clustering and constrained clustering first and then illustrate our methods through a dual-objective and dual-data sources problem.

General framework of compound clustering and constrained clustering

Suppose we have *n* datasets with *n* distance/dissimilarity measurements: D_1, D_2, …, D_n.

Compound clustering:

We formulate the overall distance as a weighted average of the individual measurements. Clusters are obtained by minimizing the overall distance D.

$$D = \lambda_1 D_1 + \lambda_2 D_2 + \cdots \lambda_n D_n \quad , where \sum_{i=1}^{n} \lambda_i = 1 \tag{4.1}$$

Constrained clustering:

Constrained clustering is an n-step algorithm as follow. In each step, we minimize D_i with the constraint that $D_j < d_j$, $j=1,2, \ldots, i-1$.

Constrained Clustering Algorithm

1. perform cluster analysis based on D_1 on all objects
2. for i = 2 to n
   Perform cluster analysis based on D_i on each cluster generated from step i, that is we minimize D_i

In a dual-objective case, that is *n=2*, compound clustering is to minimize the overall distance *D* with adjustable weighting parameter λ as follow:

$$D = \lambda D_1 + (1 - \lambda) D_2 \tag{4.2}$$

And constrained clustering is a two-step approach to minimize D_2 subject to the constraint that $D_1 \leq d$. First, we perform cluster analysis based on D_1; second, in each cluster generated in step1, we perform cluster analysis based on D_2 to obtain the final clustering results.

It can be shown that compound and constrained clustering are not equivalent. For n=2, under compound and constrained approaches, two objects can be clustered if they are close enough to satisfy:

$$D_{compound}(X_1, X_2) \leq d \Leftrightarrow X_1, X_2 \; are \; clustered;$$

$$D_1(X_1, X_2) \leq c_1 \; \& \; D_2(X_1, X_2) \leq c_2 \Leftrightarrow X_1, X_2 \; are \; clustered;$$

Then the clustering region can be shown in Figure 4.1 below:

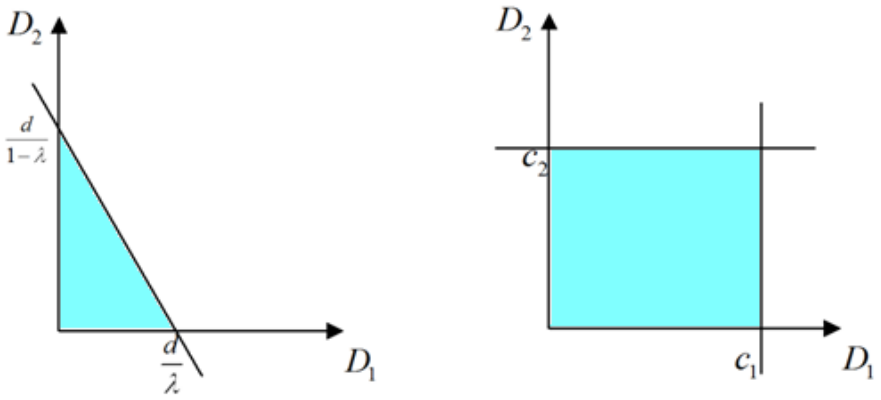


Figure 4.1. Comparison of compound and constrained clustering approaches

Visually, we can see these two approaches have different clustering regions, thus they are not equivalent. The proof can be easily extended to the case n>2. Furthermore, in order to determine the clusters, in each approach, we need to

determine two parameters when n=2, (λ,d) for the compound cluster analysis, and (c_1, c_2) for the constrained cluster analysis. In the following sections, illustrate our framework through a dual-objective and dual-data sources problem.

4.2 Our Data

The data we used is a microarray data set from the research project collaborated with Professor Josh Huang and Dr. Anirban Paul at the Cold Spring Harbor Laboratory (CSHL). It is a temporal cDNA microarray of Purkinje (PKJ) and Basket (BAS) neuron cells extracted from new born mice (Figure 4.2). The data consist of 45,000 genes from each cell. After primary extraction, about 50 PKJ cells and over 100 basket cells were retrieved with 5 time points (Figure 4.3). Our goal is to find the optimal clusters, where the genes not only share similar time course growth patterns, but also common biological functions. Thus we have two types of data to be analyzed for each cell -- one is the microarray gene expression related to time course pattern and the other is the functional data related to the biological function(s) of each gene.

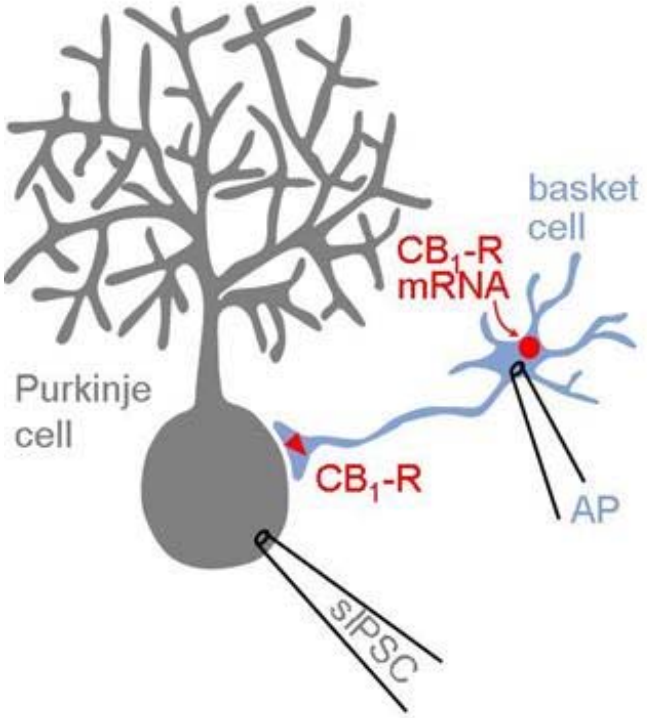


Figure 4.2 Purkinje cell and Basket cell

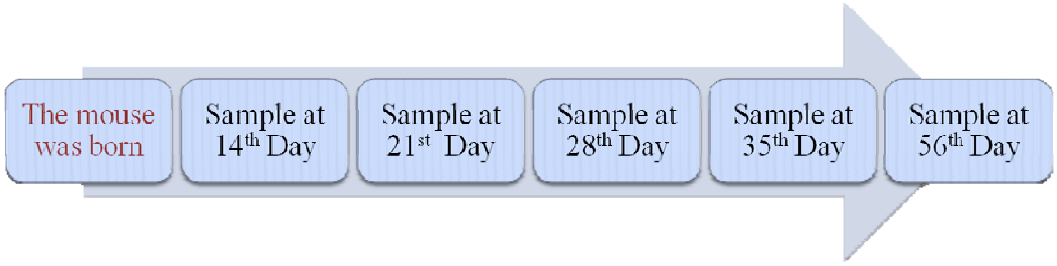


Figure 4.3 Samples are collected at 5 time points for both PKJ and BAS cells

Expression data

Data pre-processing was performed as shown below (Figure 4.4), after the time course filtering and normalization, we chose to focus on the top 1000 common genes shared by both cells.

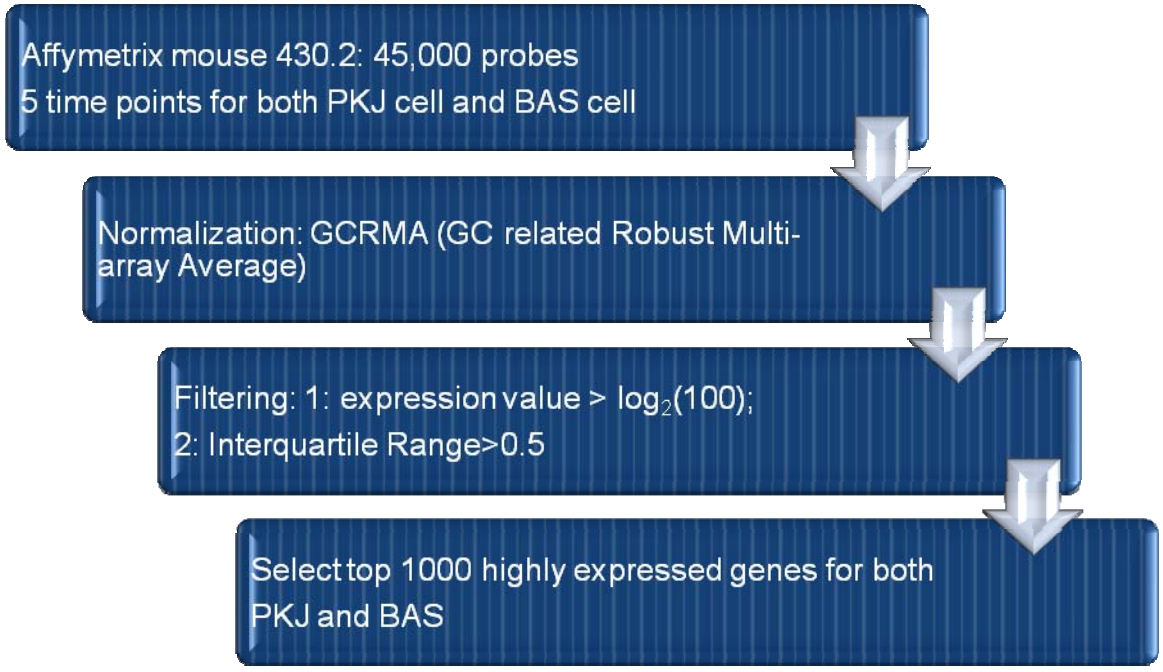


Figure 4.4 Pre-processing procedure for microarray gene expression data

Function data

Aiming to group genes based on their biological function and growth pathway information, we used the DAVID Functional Annotation Clustering (http://david.abcc.ncifcrf.gov/home.jsp) to build the function matrix. Most existing methods only incorporated one biological database. Here we input the 1000 top genes of PKJ and BAS, respectively, and integrate the GO Molecular Function, GO Biological Process, and KEGG Pathway as selection criteria to obtain 117 functional groups for each cell, PKJ or BAS (Figure 4.5).

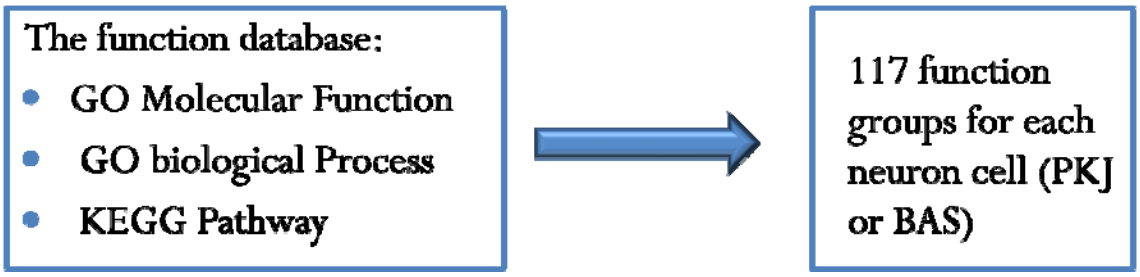


Figure 4.5 Build up the biological function groups

We subsequently built a binary function matrix [a_{ij}] for each cell. For example, for the PKJ cell, the function matrix was established as follow (Table 4.1), function groups 1- 117 as row index and the 1000 genes as column index. If gene *i* has the function group *j*, then a_{ij} is 1, otherwise a_{ij} is 0.

Table 4.1 Function matrix for the selected 1000 genes from the PKJ cell.

	Func 1	Func2	Func3	Func4	……	Func115	Func116	Func117
Gene1	1	0	0	1	……	0	0	0
Gene2	0	1	0	0	……	0	0	1
Gene3	1	0	0	0	……	0	0	0
⋮	⋮	⋮	⋮	⋮	……	⋮	⋮	⋮
Gene 999	0	0	0	0	……	1	1	0
Gene 1000	1	0	0	0	……	1	0	0

4.3 Distance Measurement

Expression distance

There are many choices of distance measures for microarray data. For expression value, researchers usually use the Euclidean distance or correlation based distance. When Euclidean distance is used, since the expression value may have

different scale at different time point or for different gene, normalization should be done first to put values on the same scale. However, Euclidean distance is based on the pair-wised distance, not the pattern similarity. As shown in Figure 2.3, the Euclidean distance between A and B is smaller than that between A and C even A and C have the same pattern, which opposite with B. Given that our aim of finding genes with similar temporal course, we find the correlation based distance to be a good choice to reflect the similarity of gene growth patterns. Because the correlation is between [-1, 1], we decided to use ½(1-correlation) as the expression distance, D_1.

Function distance

Function distance is calculated based on the binary function matrix we defined using DAVID. There are many distance measures suitable for binary data, each with different criterion and different purpose. For our goal, we want to group genes with common functions together, and we examined the following distances.

1. Euclidean / Hamming distance Although Euclidean distance and Hamming distance have distinct definition, we find they are equivalent for binary data. Hamming distance can tell us how different two vectors are. It returns the number of positions in two equal-length vectors with different values. For our project, it transpires into the number of different positions in two functional vectors of two genes.

The Euclidean distance is defined as

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_m - q_m)^2} = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2} \qquad (4.3)$$

For binary data, only when p_i and q_i are different, $(p_i - q_i)^2$ returns a value of 1, so the summation of $(p_i - q_i)^2$ equals to the number of different positions between two vectors which is exactly the Hamming distance.

2. Kappa distance We found that developers of DAVID used the Kappa score as the distance to identify the functional groups of genes (Huang, et al. 2007b). Thus we examined this as well as our second choice of functional distance measure.

The Kappa statistic is originally used in medicine and clinical research to see the agreement between categorical measurements. The Kappa statistic compares the agreement against that which might be expected by chance. For example, 29 patients are examined by two independent doctors (see Table 4.2) where 'Yes' denotes a patient being diagnosed with disease X by a doctor, and 'No' otherwise.

$$Kappa = \frac{Observed\ Agreement - Chance\ Agreement}{1 - Chance\ Agreement}$$

Table 4.2 An example of the Kappa statistic

		Doctor A		
		No	Yes	Total
Doctor B	No	10	7	17
	Yes	0	12	12
Total		10	19	29

Observed agreement = (10 + 12)/29 = 0.76
Chance agreement = 0.586 * 0.345 + 0.655 * 0.414 = 0.474
Kappa = (0.76 - 0.474)/(1 - 0.474) = 0.54

For our research, the Kappa score is defined in the same way, for two genes A and B, their Kappa score of functional agreement is as follow (Table 4.3):

Table 4.3 Kappa statistics for gene function agreement

		Gene A		
		0	1	R Total
Gene B	0	$C_{0,0}$	$C_{0,1}$	$C_{0,.}$
	1	$C_{1,0}$	$C_{1,1}$	$C_{1,.}$
C Total		$C_{.,0}$	$C_{.,1}$	$T_{A,B}$

$$\underset{Agreement}{Observed}: O_{A,B} = \frac{C_{0,0} + C_{1,1}}{T_{A,B}}$$

$$\underset{Agreement}{Chance}: A_{A,B} = \frac{C_{.,0} * C_{0,.} + C_{.,1} * C_{1,.}}{T_{A,B}^2}$$

$$Kappa = \frac{O_{A,B} - A_{A,B}}{1 - A_{A,B}}$$

Since the Kappa score ranges in [-1, 1], and for two homogeneous genes, the Kappa score is 1, we chose *(1-Kappa)/max(1-Kappa)* as the function distance measure.

4.4 Determination of Cluster Number

Since our data set is large, we steered away from simulation-based cluster number determination method initially, and instead, focused on the computational inexpensive R^2 and the largest average silhouette methods.

"Complete linkage" R^2

In fact, it is possible that genes are related in some more complicated way rather than "clusters". Hence, our clustering is probably more of a "dissection", rather than "uncovering the reality". Sarle (1996) illustrates a way to consider the R^2. Larger R^2 means the clusters can present the true structure better. But for the usual R^2, the hidden assumptions are:

1) The mean in cluster is used as the "representative component";

2) The error sum of squares is calculated in Euclidean form.

For the given data set, we adopted the "complete linkage" Hierarchical clustering method, which defines the distance between clusters by the maximum of distances between any two components. This means both assumptions are not true in our case. So we coined our own "complete linkage" R^2 as follow:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SST = \sum (X - \bar{X})^2 \Rightarrow SST \approx n * D^2$$

Because we use the "complete" linkage method, *D* here denotes the maximum distance within all *N* objects, and is used to approximate each $|x - \bar{x}|$.

$$SSE_j = \sum_{i \in I_j} (X_i - C_j)^2 \Rightarrow SSE_j \approx n_j * D_j^2$$

Here SSE_j is for cluster *j*, where n_j is the number of objects in cluster *j*, and D_j is the maximum distance within cluster *j*. Thus our "complete linkage" R^2 is as follow:

$$R^2 = 1 - \frac{\sum_j n_j * D_j^2}{n * D^2}$$

Obviously, the R^2 will achieve maximum when each gene is a cluster. In practice we go though the output and set certain value as threshold to determine the cluster numbers, as will be shown later.

The largest average silhouette (Kaufman and Rousseeuw 1990) is set as our second choice for cluster number determination. The definition of silhouette is as follow:

> Let a_i denotes the average dissimilarity between i and all other observations within the same cluster;
> Let *d(i, C)* denotes the average dissimilarity of i to all objects of *C*, where *C* is any other cluster which i doesn't belong to;
> Let b_i denotes the smallest of these *d(i, C)*;
> The silhouette of observation *i* is defined as:
>
> $$sil_i = \frac{b_i - a_i}{max\ (b_i, a_i)}$$
>
> Which is the difference between the smallest average dissimilarity between clusters and the average dissimilarity within clusters.
>
> The overall average silhouette is defined simply as the average of all sil_i:
>
> $$sil = \frac{1}{N} \sum_{i=1}^{N} sil_i$$

In practice, we also need to set the threshold when we use this parameter. Intuition/human input is needed for both threshold settings.

4.5 Heat Diagram

We developed a method to visualize the biology interpretation for each cluster, which is similar to the GO analysis. The difference is that in GO analysis, it only can indicate a few high represented function groups in each cluster, but our self-design heat diagram will allow you to see the expression pattern of each function group in each cluster. In this new visual approach, we first calculate, in each cluster, the gene frequency for each function group, and then draw the heat diagram.

This is illustrated in Figure 4.6 below. The vertical axis represents the 17 (Figure 4.6, left) and 28 (Figure 4.6, right) clusters for the BAS and the PKJ cells respectively obtained through the single-objective clustering (complete linkage hierarchical clustering) based on the microarray gene expression data (and thus the correlation based distance D_1) only. The horizontal axis 1-117 represents the 117 function groups. The last column 118 represents genes with no known function. The color bar on the right shows the gene percentage level with lighter color indicating higher percentage. Thus each "little window" in the graph reflects the percentage of gene in the corresponding cluster sharing a certain biological function or non-function group. The lighter a window is, the more genes in the corresponding cluster (indicated by vertical axis) share a common function (indicated by horizontal axis). More light colored windows in the heat diagram would indicate a high concentration of common function genes in the clusters, and thus more desirable according to our clustering objectives. However, as we can see, with almost no exception, each cluster from the single-objective clustering using the expression values only would feature genes from a diversified function groups – therefore dual-objective cluster analysis using either the compound clustering or the constrained clustering approach is necessary to produce clusters with concentrated gene functions by utilizing the gene function distance in the clustering process.

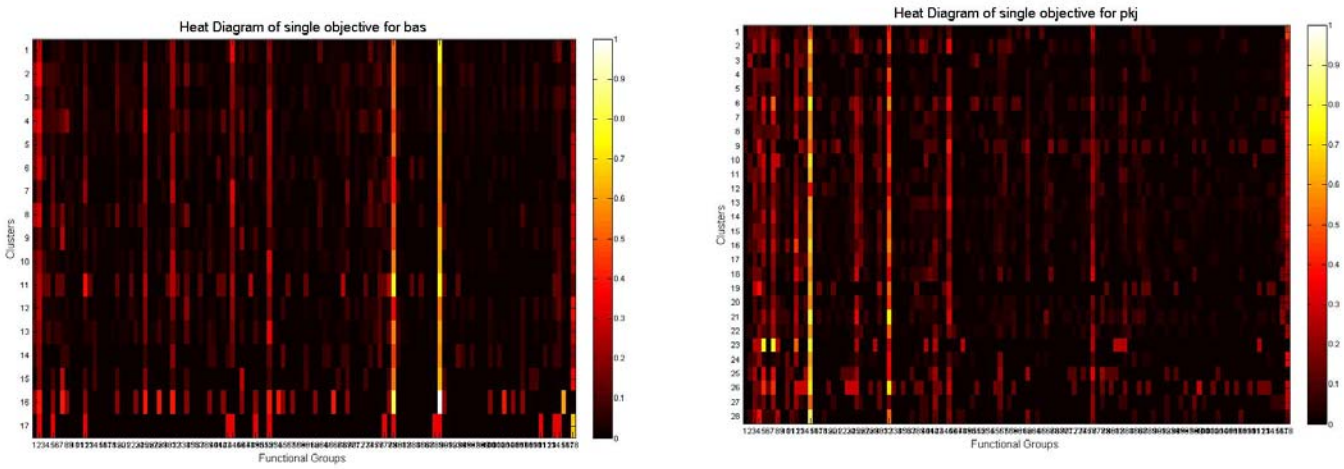


Figure 4.6 Customized cluster gene function Heat Diagram revealed that single-objective cluster analysis produced clusters with diversified gene functions for both the BAS (left) and the PKJ cells (right).

4.6 Results and Comparisons

The general procedure for compound clustering and constrained clustering is shown in Figure 4.7. In analyzing the neuron cell data from CSHL, we used correlation-based distance as the gene expression distance D_1, and tried two types of functional distance measure D_2, the Hamming/Euclidean distance and the Kappa statistic based distance. We also examined two approaches, the "Complete linkage" R^2 and largest average silhouette, to determine the parameters for compound clustering. However, it is difficult to determine parameters for constrained clustering using either the "Complete linkage" R^2 and largest average silhouette, which we will discuss more in Chapter 8. In this section, we performed the compound clustering using the hierarchical clustering algorithm. We first show the result from our compound cluster analysis in one example and then compared with the traditional cluster analysis utilized only gene expression data as well as the effect of two different functional distance measures and two statistics for cluster number determination.

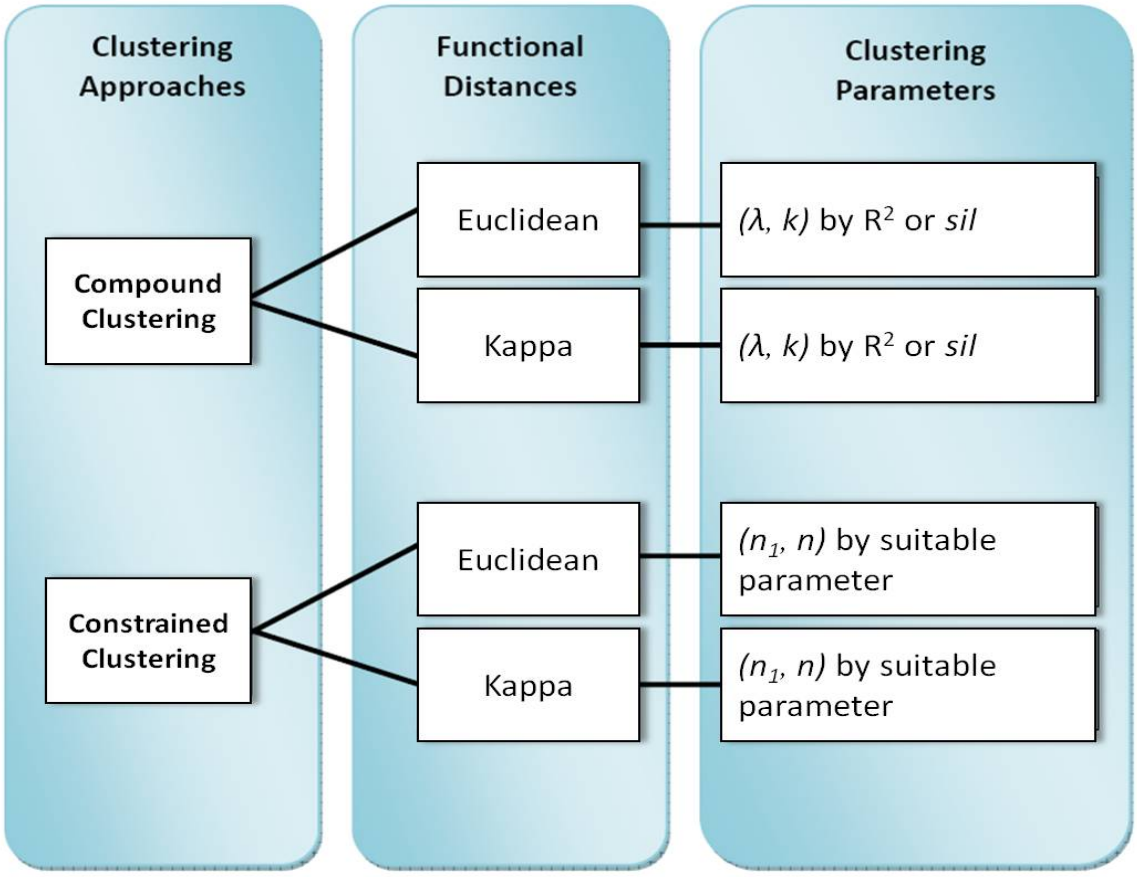
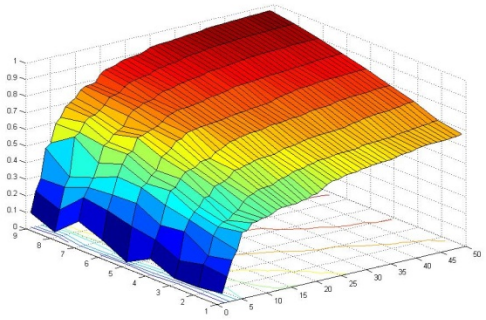


Figure 4.7 General procedure for compound clustering and constrained clustering

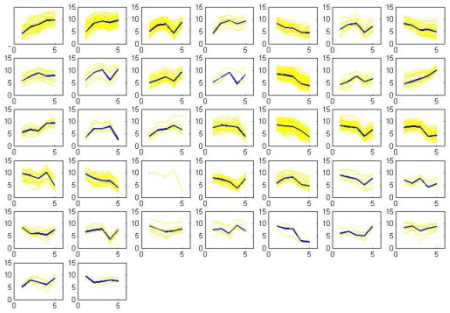
4.6.1 Results of compound clustering

Here we first give an example of the compound clustering. For compound clustering, we need to determine *(λ, n)* in order to get the optimal result, where *λ* is the compound clustering weight, and *n* is the total cluster number. In Figure 4.8, they are the graphical outputs for the BAS cell with the Euclidean distance as D_2, and R^2 as the cluster number determination parameter.

a.



b.



c.

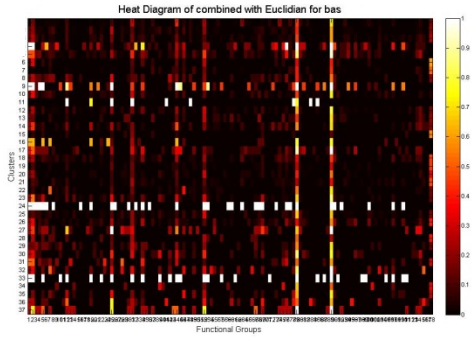


Figure 4.8 Compound clustering result for BAS cell with Euclidean distance and R^2.

Figure 4.8(a) is the 3 dimensional plot of R^2 against *(λ, k)* which is used to decide the cluster number *k* and parameters *λ*. We can see that the R^2 increase when *λ* goes up or *k* goes up. As we know, a large cluster number renders the result less useful and meaningful; however, a higher R^2 is always preferred. So determining the cluster number *k* and parameter *λ* is actually to make the trade-off between R^2 and cluster number. In order to find the best combination, we decide to set $R^2 \geq 0.9$, and search for the λ that will yield the smallest number of clusters. The result is *λ*=0.7 with a corresponding cluster number of 37. Figure 4.8(b) reveals the temporal gene expression pattern for these 37 clusters while Figure 4.8(c) demonstrates the cluster gene function patterns using our customized function heat diagram. We can see that the temporal patterns are distinct from cluster to cluster, and the gene function groups are concentrated within clusters.

4.6.2 Comparison

We performed compound cluster analysis on both BAS and PKJ cells with two functional distance and two statistics to determine cluster number and generated four

sets of results with different setting. For each set of results, *(λ, k)* are determined as the way we described above.

We summarized the results for the four sets of compound cluster analysis in Table 4.4. We found that the R^2 criterion yields consistent *λ* for both the BAS and the PKJ cells. It also features that with around 35 clusters, it yields high R^2 values for both function distances. The largest average silhouette method, however, produces different *λ*s for different cells, and furthermore, the criterion value is negative which is hard to explain. As we introduced before, Silhouette is defined as the difference between average dissimilarity between clusters and the average dissimilarity within cluster. Therefore, it should be positive because the dissimilarity within cluster should be smaller than the dissimilarity between clusters. The negative Silhouette in our case may because of the weighted overall distance we employed, which makes Silhouette is not feasible for our program. In terms of interpretability and stability, R^2 criterion is superior for this data set than the largest average silhouette criterion for cluster number determination.

Table 4.4 Summary of results from the compound cluster analysis

Cluster Number Determination	D2	Cell Type	λ	Cluster Number	Cluster Number Criterion Threshold
R^2	Euclidean	BAS	0.7	37	$R^2 > 0.90$
		PKJ	0.7	35	
	Kappa	BAS	0.8	30	$R^2 > 0.95$
		PKJ	0.8	34	
Silhouette	Euclidean	BAS	0.4	33	Silhouette > -0.3
		PKJ	0.7	35	
	Kappa	BAS	0.3	35	Silhouette > -0.3
		PKJ	0.2	31	

We subsequently compared the ability of finding biological meaningful clusters between our compound clustering incorporating biological information to the original single-objective hierarchical clustering method. Both the Euclidean distance and the Kappa distance are used with R^2 to determine the cluster number. Heat diagram is used to see which method could give more biological meaningful clusters with more light

colored windows. As discussed in Section 4.5, light colored window in heat diagram indicates the common functions shared by the genes in each cluster. Therefore, more light colored windows in heat diagram, more shared functions in the resulted clusters. Figure 4.9 (B1, P1) feature the Euclidean distance while Figure 4.9 (B2,P2) showcase the Kappa distance and Figure 4.9 (B3,P3) represent the original hierarchical clustering method. We found that with Euclidean distance as functional distance, the algorithm yields much better heat diagrams with more light colored windows.

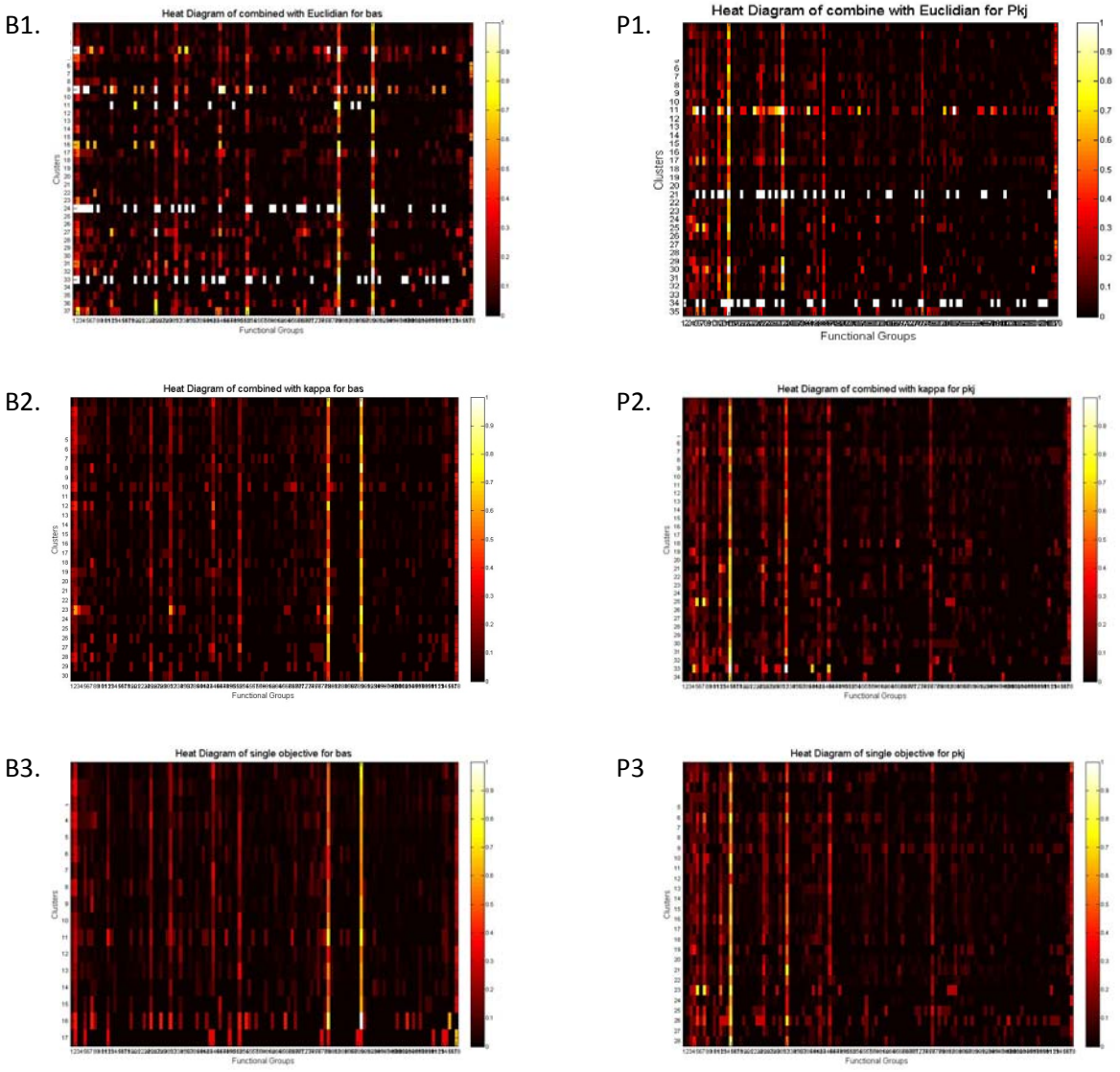


Figure 4.9 Heat diagram of clusters and gene function groups under different gene function distance measure as well as original hierarchical clustering when R^2 is adopted as the cluster number determination parameter. B: results from BAS cell; P: results from PKJ cell. 1: Euclidean distance as

functional distance; 2: Kappa distance as functional distance; 3: original hierarchical clustering with no functional distance.

In summary, we performed compound clustering on microarray data to illustrate our method. We used two functional distance and two statistics to determine the appropriate parameter *(λ, k)* for compound clustering. From this application, we can conclude that Euclidean distance as functional distance with our newly proposed "Complete Linkage" R^2 is the best combination for compound clustering to generate most biological meaningful cluster result. With the best combination, we find the meaningful clusters with enriched biological functions as shown below (Table 4.5).

Table 4.5 Selected clusters with related biological functions in BAS cell using Euclidean distance and "Complete Linkage" R^2

Cluster #	# of Genes	Related Biological Function Groups
2	69	Lipid Transport, Homeostasis
3	85	Sphingoid Metabolic Process
5	34	Calcium-mediated Signaling,
10	36	Regulation of Cell Morphogenesis
12	102	Peptide Binding
14	23	Serine-type Enzyme
18	31	Neuronal Structure Regulation
19	90	Urogenital System development
20	86	Multicellular Organism Growth, GTPase Binding
21	54	Peroxidase activity

Part II. Weighted K-means Clustering

Chapter 5

Weighted K-means Clustering

5.1 Current Issues in K-means Clustering

K-means clustering is the most popular partitional clustering method. The algorithm is straight forward and easy to implement. With time complexity of *O(n)*, k-means clustering is capable and favorable to run on high dimensional data structure. However, each method has its own limitations, so does the k-means clustering. It is well known that k-means clustering may end up with local optimal solution, depending on the starting values used. Determining the cluster number k in advance is also very difficult and no theoretical solution has been reported on this. At last, researchers have recognized that not all variables or features make the same contributions in clustering. So how to select and use those variables is another important issue in this field.

5.1.1 Initialization and local optimal

K-means clustering starts from a random generated set of k points as initial k cluster centers and then, ideally, iteratively relocate to k true cluster centers which minimize the with-in cluster sums of squares. However, it has long been known that the algorithm may not converge to the global minimum depending on the initial k points.

In order to find the global optimum result, initialization is the key. Various methods have been published in the literature to deal with this problem. The first method was proposed by Astrahan in 1970. He assigned a density to each point based on the k-nearest neighbor. Then the points, within a specified distance from the selected initial points, had very low probability to be chosen into the initial set. A similar method is currently implemented in the PROC FASTCLUS procedure in SAS (SAS 2004). Hajnal and Loosveldt (2000) argued that the initialization employed in SAS is better compared

to random initialization for k-means clustering. At the same time, many literature suggested using the cluster center obtained from another clustering algorithm, most likely hierarchical clustering as the initial cluster centers, instead of randomization (Milligan 1980; Punj and Stewart 1983; Arabie and Hubert 1994; Huberty, et al. 1997). However, the results from hierarchical clustering may not be the ideal case and it is cumbersome to perform additional k-means clustering on top of the hierarchical clustering. Bradley and Fayyad (1998), on the other hand, proposed a bootstrap algorithm to determine the initial points for k-means clustering.

However, all the methods mentioned above required extra computation time, which compromise the time efficiency of k-means clustering. The most popular solution for initialization problem is to perform k-means clustering several time with different initial sets first (Makarenkov and Legendre 2001) and then select the best one. Steinley (2003) compared several initialization methods and confirmed the best one is to use multiple initial sets. Even with initialization issue, it has been shown that k-means clustering consistently performs reasonable well in recovering data structure (Dimitriadou, et al. 2002; Steinley 2003). In this thesis, we performed k-means clustering and weighted K-means clustering with multiple initial sets to achieve the best performance.

5.1.2 What is the appropriate cluster number K

K-means clustering requires determining the cluster number K before performing the algorithm. However, determining the appropriate cluster number is one of the most difficult problems in cluster analysis. We have introduced the majority of methods in determining cluster number in Section 3.4. Most of the methods introduced there are also widely used in k-means clustering. However, most the methods require to run k-means clustering several times and select the best one among them, which needs additional computation time. It has just been proven in a recent paper that performing the k-means clustering with flexibility on K is NP-hard even in 2-dimensional space (Mahajan, et al. 2009). In this thesis, we only examine the performance of our proposed method with the correct cluster number. Determining cluster number K for k-means clustering is not the focus of this work.

5.1.3 How to use the variables

After determining the cluster number K, data needs to be preprocessed before performing k-means algorithms. Common steps include variable standardization, selection, reduction and variable weighting.

Variable standardization is usually conducted before performing k-means cluster analysis. Milligan and Cooper (1988) investigated eight different standardization methods which is the most comprehensive review on variable standardization in clustering. Dillon, Mulani and Frederick (1989) showed standardized data could result in significant difference in cluster results compared to the data without standardization which illustrated the importance of variable standardization. Vesanto (2001) suggested that z-score, Equation 5.1, should be the first choice in variable standardization because of the easy interpretability. Steinley (2004) repeated Milligan and Cooper's study focusing on k-means clustering and recommended standardization by range, equation 5.2, instead of z-scores.

$$z-score: \quad z_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_j} \tag{5.1}$$

$$standardization\ by\ range: \quad z_{ij} = \frac{x_{ij}}{max(x_{.j}) - min(x_{.j})} \tag{5.2}$$

Variable selection is another important step before performing cluster analysis, because of the well known facts that all variables are not equally important, and furthermore, some are highly correlated and thereby redundancy results when selected collectively. Compared to supervised learning, automated variable selection in unsupervised learning is much more difficult and challenging due to lack of class information. Friedman and Rubin (1967) first discussed the influence of variable selection in clustering by assigning a binary 0 or 1 variable weights. Fowlkes and colleagues (1988) discussed three subset selection methods: forward selection, backward selection and stepwise selection, just as the subset selection methods in linear regression, to find the best subset in hierarchical clustering and note that these methods can be extended to K-means clustering. Carmone and co-workers (1999) proposed a variable selection technique called HINoV based on the adjusted Rand

index (Hubert and Arabie 1985), which is used for cluster validation and cluster number determination. Then Brusco and Cradit (2001) presented several limitations with HINoV and developed a better variable selection method for K-means clustering named VS-KM. A recent summary and comparison of different variable selection methods in k-means clustering can be found in Steinley and Bruce (2008).

Variable reduction is used when dealing with high dimensional data with correlated variables. Common data reduction techniques are combined with k-means clustering. Principal component analysis (PCA) was the first (Barker 1976) and most popular one used in k-means clustering. In this method, PCA is performed first on the data, and then the principal components with a corresponding eigenvalue greater than one will be used to perform k-means clustering. This technique is frequently used in high dimensional data analysis (De Backer and Scheunders 1999; Ben-Hur, et al. 2002). But some researchers criticized that the first few components may not guarantee a subspace that provides enough information about the true structure in the data (Arabie and Hubert 1994; De Soete and Carroll 1994). Instead, De Soete and Carroll (1994) proposed a method to minimize the sums of squares between the full dimensional data and the low-dimensional cluster centers defined by multidimensional scaling (MDS). Some other methods have been developed using MDS (Van Buuren and Heiser 1989; Heiser and Groenen 1997; Vichi and Kiers 2001). However, a recent comparison between k-mean clustering and PCA has theoretically proved that “Cluster centroid subspace is spanned by the first k-1 principal directions” (Ding and He 2004).

In addition, variable weighting, which can be thought as a generalization of variable selection, is also a very interesting topic and becomes popular recently. We will discuss weighted cluster analysis in next section.

5.2 Existing Weighted K-means Methods

Weighted Cluster analysis is an extension of the classical clustering analysis in which a set of nonnegative weights are assigned to all the variables and then cluster

analysis is performed on the set of weighted variables. Variable weighting, which increases or decreases the influence of variables by different weighting scheme, can be thought as a generalization of variable selection, in which the weights are restricted to either 1 or 0 (Wettschereck, et al. 1997). Steinley and Brusco (2008) recently proposed a variance-based index which can be applied on both variable weighting and variable selection. It has been shown that in real application, by selecting appropriate variable weights, k-means clustering performs better and achieves a more meaningful and interpretable results (Tseng 2007; Shen, et al. 2010).

The key step in weighted k-means clustering is to estimate the optimal variable weights, on which the weighted clustering will be performed. There are already some studies conducted in variable weighting for clustering analysis. In estimating variable weights, we usually restrict that all the weights sum up to *m*, however, in few literature, the weights are restricted with a sum of 1. Variables with zero weight are actually excluded from the subsequent analysis. Then variables with a weight above 1 (or 1/*m*, if with restricted sum of 1) are considered more informative about the underlying data structure, while variables with a weight less than 1 or 1/*m* are considered less informative. There are already some studies conducted in variable weighting for clustering analysis. Friedman and Rubin (1967) first introduced the concept of variable weighting in the discussion about variable selection with variable weights of 0 and 1. Generally, weighted Euclidean distance in k-means clustering is defined as in 5.3. However, variable weights were defined differently in the past Literature, which can be roughly summarized into three categories.

$$d_{ik}^2(w) = \sum_{j=1}^{m} w_j^2 (x_{ij} - x_{kj})^2 \tag{5.3}$$

Feature-based weighting

In feature-based weighting, researchers focused on balancing the influences of multiple groups of features by assigning weights on each group of variables. Desarbo and colleagues (1984) proposed SYNCLUS model with a two-level weighting scheme on each group of variables as well as each variable individually. Then a two-step

iteration algorithm was employed to estimate variable weighting on both variable level and group level. Modha and Spangler in 2003 improved Desarbo's method by considering the weights only on variable level and minimizing the ratio of with-in cluster sum-square over between-cluster distortion.

Distance-based weighting

In distance-based weighting, variable weighting were used to make sure the weighted distance, defined in 5.3 is geometrically well-defined. De Soete (1986) introduced the method for optimal variable weighting by minimizing the following objective functions so that the weighted distance satisfies the ultrametric inequality (5.4) and additive-tree inequality (5.5).

$$L_U(w) = \frac{\sum_{\Omega_U}(d_{ik} - d_{jk})^2}{\sum_{i<j} d_{ij}^2}, \qquad \Omega_U = \{(i,j,k)|d_{ij} \leq min\ (d_{ik}, d_{jk})\} \tag{5.4}$$

$$L_A(w) = \frac{\sum_{\Omega_A}\left(d_{ik} + d_{jl} - d_{il} - d_{jk}\right)^2}{\sum_{i<j} d_{ij}^2},$$
$$\Omega_A = \begin{Bmatrix}(i,j,k,l)|d_{ij} + d_{kl} \leq \min\ (d_{ik} + d_{jl}, d_{il} \\ +d_{jk})\end{Bmatrix} \tag{5.5}$$

Then Makarenkov and Legendre (2001) extended De Soete's idea to estimate variable weights in k-means clustering by minimizing the following weighted sums of squared Euclidean distance in 5.6 along with 5.4 and 5.5 using Polak-Ribière optimization methods.

$$L_P(w) = \sum_{k=1}^{k}\left[\sum_{i,j=1}^{n_k} d_{ij}^2\right]/n_k \tag{5.6}$$

Objective-based weighting

It has been reported (Gnanadesikan, et al. 1995) that both SYNCLUS and De Soete's weighting method performed poorly in real data because the way the variable weights were defined may not directly guarantee an optimal clustering performance. Recently, objective-based weighting method, which estimates the optimal variable weights by minimizing the weighted objective function in k-means clustering:

$$\sum_{g=1}^{k}\sum_{i\in I_g}\sum_{j=1}^{m} f(w_j) * (z_{ij} - c_{gj})^2 \tag{5.7}$$

Huang's group (2005) proposed a three-step iteration method for optimal variable weight by minimizing function 5.8 with extra scale parameter *u* weighting parameter *β*. This method results in balanced cluster size and high efficiency on large data sets but lower interpretability on the variable weighting. Several variants have been proposed (Tseng 2007; Tsai and Chiu 2008; Shen, et al. 2010).

$$\sum_{g=1}^{k}\sum_{i\in I_g}\sum_{j=1}^{m} u_{i,g} w_j^{\beta} (z_{ij} - c_{gj})^2 \tag{5.8}$$

However, one big drawback of all the studies mentioned above is instability on variable weighting. The resulted optimal variable weighting is very sensitive with the data. Most recently, Huh and Lim (2009) claimed the instability problem was solved with a new objective function to minimize the weighted sums of squares with an extra penalty term as follow:

$$\sum_{g=1}^{k}\sum_{i\in I_g}\sum_{j=1}^{m} \frac{w_j (z_{ij} - c_{gj})^2}{n-1} + \alpha \sum_{j=1}^{m} \frac{(w_j - 1)^2}{m-1} \tag{5.9}$$

They used process optimization in response surface methodology to estimate the optimal variable weighting and resulted in stable variable weighting by adjusting the penalty parameter *α*. However, their method performed poorly with large data set. More important, Nelder-Mead optimization algorithm (Nelder and Mead 1965) they employed cost a lot on running time with increasing dimensionality of data and is not global optimal guaranteed.

In this thesis work, we theoretically derive the optimal variable weights based on 5.9 and propose a new algorithm to solve it. The theoretical part is illustrated in chapter 6 while the numerical examples and comparisons are presented in Chapter 7.

Chapter 6

Proposed Method and Algorithm

As we discussed in Chapter 5, variable weighting in k-means clustering, as a generalization of variable selection, has shown good promise and garnered increased attention in recent years. Recently, Huh and Lim (2009) proposed a penalized objective function that has achieved stable variable weights when applied to low dimensional data. However, their method performs poorly for high-dimensional data as shown in their paper. Furthermore, the optimization algorithm they utilized may result in local rather than global optimality. In this thesis work, we use the same penalized objective function Huh and Lim has proposed. However, instead of relying on numerical optimization directly, we first derive the theoretical solution for global optimal variable weights and then implement a companion numerical algorithm to compute the desired weights. Our approach is shown to generate more stable variable weights for high dimensional data – and thereby achieve better clustering accuracy.

6.1 Close-form Solution for Variable Weights

The objective function 5.9 can be written as a function of weights $(w_1, w_2, \dots, w_m)$*:*

$$\sum_{g=1}^{k} w_j \sum_{i \in I_g} \sum_{j=1}^{m} \frac{(z_{ij} - c_{gj})^2}{n-1} + \alpha \sum_{j=1}^{m} \frac{(w_j - 1)^2}{m-1} \tag{6.1}$$

The coefficient of weight w_j in the first term is "Within-cluster mean squares on *jth* variable"(*j*WCMS). Assuming the true cluster centers $C_g,\ g = 1,2,\dots,k$ are known, we can denote *j*WCMS as β_j. Without losing generality, we can assume $\beta_1 \le \beta_2 \le \cdots \le \beta_m$.

Then when α is given, the optimal variable weighting is the solution of minimizing the following quadratic function with inequality constraints:

$$Minimize:\ f(w;\alpha) = \sum_{g=1}^{k} \beta_j w_j + \alpha \sum_{j=1}^{m} \frac{\left(w_j - 1\right)^2}{m-1}$$

$$Subject\ to:\ \sum_{g=1}^{k} w_j = m; \tag{6.2}$$

$$w_j > 0, j = 1,2,\dots,m.$$

Harold W. Kuhn, Albert W. Tucker and William Karush generalized the method of Lagrange multipliers to solve nonlinear programming with inequality constraints (Karush 1939; Kuhn and Tucker 1951). In their approach, the optimal solution of a nonlinear programming problem must satisfy a set of conditions, which are called Karush-Kuhn-Tucher (KKT) conditions. Here we solve above minimization problem using KKT conditions and subsequently find the optimal weighting.

First of all, the Lagrange function of (6.2) is as follow:

$$L(w,\lambda,\mu;\alpha) = \sum_{g=1}^{k} \beta_j w_j + \alpha \sum_{j=1}^{m} \frac{\left(w_j - 1\right)^2}{m-1} + \lambda\left(\sum_{g=1}^{k} w_j - m\right) + \sum_{g=1}^{k} \mu_j w_j \tag{6.3}$$

Then according to KKT conditions, the optimal weights $(w_1, w_2, \dots, w_m)$ must satisfy the following:

$$\frac{\partial L}{\partial w_j} = \beta_j + \frac{2\alpha}{m-1}\left(w_j - 1\right) + \lambda + \mu_j = 0, \qquad j = 1,2,\dots,m \tag{6.4}$$

$$\sum_{g=1}^{k} w_j - m = 0 \tag{6.5}$$

$$\mu_j w_j = 0, \qquad j = 1,2,\dots,m \tag{6.6}$$

$$w_j > 0, \qquad j = 1,2,\dots,m \tag{6.7}$$

Solving the above equation system, we can find the optimal variable weighting.

Before solving for the optimal variable weighting, we first prove the following proposition.

***Proposition* 6.1**

When minimum of the following function is reached, $w_i > w_j$ if and only if $\beta_i < \beta_j$.

Proof: *Assume* $\exists\ w_i < w_j$ *and* $\beta_i < \beta_j$, *when* $f(w;\alpha)$ *is minimum*;

Then $\beta_i w_i + \beta_j w_j - \beta_i w_j - \beta_j w_i = (\beta_i - \beta_j)(w_i - w_j) > 0 \Leftrightarrow \beta_i w_i + \beta_j w_j > \beta_i w_j + \beta_j w_i$;

Therefore, $f(w;\alpha)$ can be further minimized by switching (w_i, w_j),

which is contradicted with minimum of $f(w;\alpha)$

Now we start to solve the equation system. From (6.6), for each *j*, either μ_j or w_j must be zero. Assuming there are only t variables have nonzero weights, based on (6.7) and Proposition 1, we have:

$$w_1 \geq w_2 \geq \cdots \geq w_t > 0 = w_{t+1} = \cdots = w_m \tag{6.8}$$

$$\mu_1 = \mu_2 = \cdots = \mu_t = 0 \tag{6.9}$$

Then substitute those zeros into equation 6.4 and 6.5, we can get the solution with t nonzero weights:

$$\begin{cases} \mathrm{w(j;\alpha,t)} = & \frac{\mathrm{m}}{\mathrm{t}} + \frac{(\overline{\beta}_t - \beta_j)(\mathrm{m}-1)}{2\alpha} & \mathrm{j} \leq \mathrm{t} \\ \mathrm{w(j;\alpha,t)} = & 0 & \mathrm{j} > t \end{cases} \tag{6.10}$$

where, $\overline{\beta}_t = \frac{\sum_{i=1}^{t}\beta_i}{t}$, $\lambda(\alpha,\mathrm{t}) = -\overline{\beta}_t$.

In order to find the optimal variable weights, we then need to decide t, number of nonzero variable weights. Assuming t* is the true value of t, from equation 6.8,

$$w_{t^*} > 0 \tag{6.11}$$

In the other hand, if we mis-specify t= t*+1, equation 6.8 will be violated, which yields:

$$w_{t^*+1} \leq 0 \tag{6.12}$$

Then from equation 6.11, we get:

$$w_t = \frac{m}{t} + \frac{(\overline{\beta}_t - \beta_j)(m-1)}{2\alpha} > 0 \Rightarrow \alpha > \frac{t\left(\beta_j - \overline{\beta}_t\right)(m-1)}{2m} \tag{6.13}$$

Then denote

$$g(t) = \frac{t\left(\beta_j - \overline{\beta}_t\right)(m-1)}{2m}$$

All possible t for a given α is a set which satisfies:

$$T(\alpha) = \{t | g(t) < \alpha \leq g(t+1)\} \tag{6.14}$$

The optimal t is:

$$t_{opt} = \underset{t \in T(\alpha)}{argmin}\, f(w;\alpha) \tag{6.15}$$

Therefore, replacing t by t_{opt} in solution 6.10, we finally get the optimal variable weighting:

$$\begin{cases} w(j;\alpha) = \dfrac{m}{t_{opt}} + \dfrac{\left(\overline{\beta}_{t_{opt}} - \beta_j\right)(m-1)}{2\alpha} & j \leq t_{opt} \\ w(j;\alpha) = 0 & j > t_{opt} \end{cases} \tag{6.16}$$

6.2 Iteration Algorithm

In 6.1, we derived the close-form solution for variable weighting in k-means clustering when β (within-cluster mean squares on each variable) is known for all variables. In reality, however, there is no way to know the actual βs unless you know the clustering partition. To solve this, we propose an EM-like two-step algorithm to estimate βs and α iteratively and therefore find the optimal variable weighting. In step 1, we update the variable weighting using βs and α; then in step 2, we calculate new βs by performing weighted k-means clustering on weighted variables and subsequently select α. These two steps are repeated until βs converge.

Iteration Algorithm

Input Initial *β* estimation, penalty parameter *α* and standardized data matrix *Z*

Repeat

3. calculate optimal variable weight vector $(w_1, w_2, \dots, w_m)$ using 6.16 with penalty parameter *α* and *β*s;
4. Run k-means on weighted variable $Z^* = Z * D$, where D is a diagonal matrix with $Diag(D) = (\sqrt{w_1}, \sqrt{w_2}, \dots, \sqrt{w_m})$ and calculate within-cluster mean squares on each variable as *β*s and updated penalty parameter *α*;

Until *β*s converge

6.3 Initial β Estimation

After simple derivations in 6.17, we can easily show the following linear relationship between the overall within cluster sums of squares on weighted variables Z^* and the within cluster mean squares on each original variable, *β*s.

$$\sum_{g=1}^{k} \beta_j w_j = \sum_{g=1}^{k} \sum_{i \in I_g} \sum_{j=1}^{m} \frac{w_j (z_{ij} - c_{gj})^2}{n-1} = \frac{1}{n-1} \sum_{g=1}^{k} \sum_{i \in I_g} \sum_{j=1}^{m} (z_{ij}\sqrt{w_j} - c_{gj}\sqrt{w_j})^2 = \frac{1}{n-1} \sum_{g=1}^{k} \sum_{i \in I_g} \sum_{j=1}^{m} (z_{ij}^* - c_{gj}^*)^2 \quad (6.17)$$

Given the constraint that all the variable weights w_j sum up to *m*, we formulate the following canonical mixture linear model with *β*s as coefficients and y is the within cluster mean squares on weighted variable Z^*.

$$y = \sum_{g=1}^{k} \beta_j w_j + \varepsilon \quad (6.18)$$

To estimate initial *β*s, we apply a {*m, 2*} simplex lattice design with center point to generate initial variable weighting and estimate *β*s afterward. Generally, a {*m, p*} simplex lattice design generates a set of m-dimensional points $(x_1, x_2, \dots, x_m)$ such that each component can take the *p+1* equally spaced values from 0 to 1, that is, $x_i = 0, 1/p\,, 2/p\,, \dots, 1; for\ i = 1,2, \dots, m$ and the sum of all the component equal to 1. Graphically, it consists of all *m* vertices and *p* equal-division-points on $\binom{m}{2}$ edges of *m-1* dimensional simplex. For example, a {3, 2}simplex lattice design (Cornell 2002) with center point consists of the following 6 points (Fig 6.1), which are also 3 vertices, midpoints of 3 edges of 2-simplex (the equilateral triangle) and the center.

$$\{1,0,0\}, \{0,1,0\}, \{1,0,0\}, \{0, 1/2\,, 1/2\}, \{1/2\,, 0, 1/2\,, \}, \{1/2\,, 1/2\,, 0\}, \{0,0,0\}$$

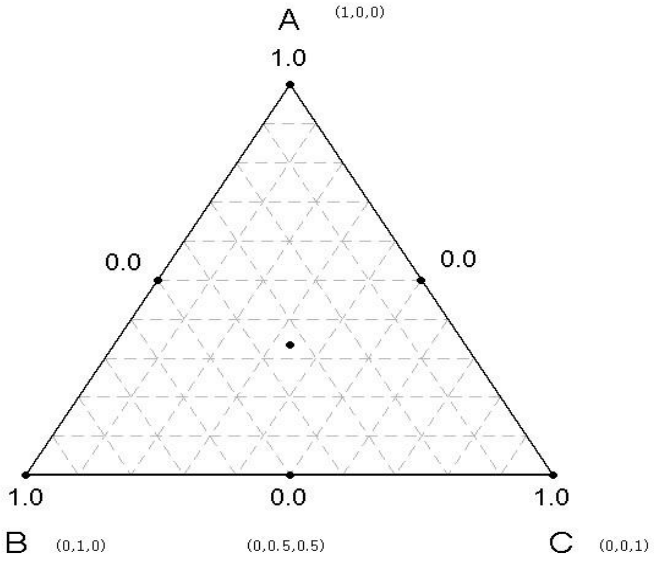


Figure 6.1 {3, 2} simplex lattice design with center point

So back to our problem, we generate a {*m, 2*} simplex lattice design as a set of vector $P = (p_1, p_2, \dots, p_m)$. Then for each design point *P*, we run the k-means clustering on weighted variable with weighting $W = m * P$ and calculate the overall within cluster sum of squares and then the response variable y in model 6.18. After this, we fit the linear model and calculate the least square estimator as initial *β*s.

6.4 Selection of the Penalty Parameter α

In optimal solution 6.16, the optimal variable weighting is a function of α. Nonnegative parameter α here is the penalty for heterogeneity in variable weighting, which is also a tuning parameter to stabilize the optimal weights (Huh and Lim 2009). Considering the extreme case of *α=0*, the objective function (6.2) is minimized when the variable with smallest β has weighting m and all other variables has zero weight, that is, $w_1 = m, w_2 = \cdots = w_m = 0$. Then by gradually increasing *α*, the penalty for heterogeneity increases as well and all variable weights move towards 1. Therefore, choosing an appropriate value for α is critical.

Huh and Lim proposed a two-step split sample method to select *α* in order to have stable variable weighting. In their method, a range was determined for *α* in the first step so that all variables have stabilized weights within that desired range; then the whole sample is split into two to check the stability. However, determining a desirable range is very subjective with personal bias and could be very hard when there are lots of variables, as shown in their paper. Also, splitting sample could be very tricky in terms of remaining the similar structure in each split sample. If one sample mainly contains two clusters and another sample contains subjects from remaining clusters, it doesn't make sense to expect that both samples yield similar variable weightings. Furthermore, their method mainly focuses on the stability of weighting while may losing the optimality.

Here we propose a method of determining *α*, based on variable selection, in order to generate optimal clustering result. Recall in the derivation (6.15) and (6.16), α determines not only the variable weighting but also the number of nonzero variable weights t. However, α doesn't determine clustering assignment. Instead, the number of nonzero variable weights *t*, which is decided by *α*, determines clustering assignment directly. That is, if we just change α within the range $(g(t), g(t+1)]$ without changing t, only optimal weighting is changed but not clustering assignment. We prove this as the following proposition.

***Proposition* 6.2**

$$z_0 = \{z_{0j}\} \in C_{g_0}, for\ all\ \alpha \in (g(t), g(t+1)], if\ \exists \alpha^* \in (g(t), g(t+1)], so\ that\ z_0 = \{z_{0j}\} \in C_{g_0}.$$

Proof: *First, we define the weighted squared-distance between object* z_0 *and cluster center* C_{g0} *as follow:*

$$D_\alpha(z_0, g_0) = \sum_{j=1}^{m} (z_{0j}\sqrt{w_j(\alpha)} - c_{g_0 j}\sqrt{w_j(\alpha)})^2 \tag{6.19}$$

Then define function F as the difference between two weighted squared-distances:

$$F_\alpha(z_0, g_0, g_i) = D_\alpha(z_0, g_0) - D_\alpha(z_0, g_i) \tag{6.20}$$

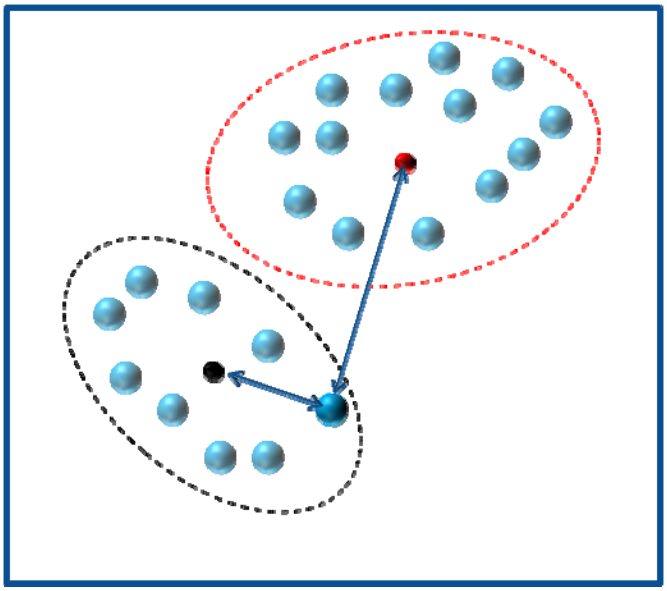


Figure 6.2 Illustration of k-means clustering

As shown in Figure 6.2, in k-means clustering, one object is always assigned to the nearest cluster with smallest distance to the cluster center. That is,

$$\begin{gathered} z_0 = \{z_{0j}\} \in C_{g_0} \Leftrightarrow \\ D_\alpha(z_0, g_0) < D_\alpha(z_0, g_i) for\ \forall i \neq 0 \\ \Leftrightarrow F_\alpha(z_0, g_0, g_i) < 0 \end{gathered} \tag{6.21}$$

Therefore, Proposition 6.2 is mathematically equivalent to the following statement:

$$F_{z_0,g_0,g_i}(\alpha) < 0 \; for \; all \; \alpha \in (g(t), g(t+1)], \\ if \; \exists \alpha^* \in (g(t), g(t+1)], s.t. F_{z_0,g_0,g_i}(\alpha^*) < 0 \, . \tag{6.22}$$

Thus, we can prove this one instead. First, we can show $F_{z_0,g_0,g_i}(\alpha)$ is actually a Hyperbolic function of α with location parameter H1 and scale parameter H2.

$$F_{z_0,g_0,g_i}(\alpha) = D_\alpha(z_0, g_0) - D_\alpha(z_0, g_i)$$

$$= \sum_{j=1}^{m} \left\{ \left(c_{g_i j} \sqrt{w_j(\alpha)} - c_{g_0 j} \sqrt{w_j(\alpha)} \right) \left(2 z_{0j} \sqrt{w_j(\alpha)} - c_{g_0 j} \sqrt{w_j(\alpha)} - c_{g_i j} \sqrt{w_j(\alpha)} \right) \right\}$$

$$= \sum_{j=1}^{m} \left\{ w_j(\alpha) \left(c_{g_i j} - c_{g_0 j} \right) \left(2 z_{0j} - c_{g_0 j} - c_{g_i j} \right) \right\}$$

$$= \sum_{j=1}^{t_{opt}} \left[\frac{m}{t_{opt}} + \frac{\left(\overline{\beta}_{t_{opt}} - \beta_j \right)(m-1)}{2\alpha} \right] \left\{ \left(c_{g_i j} - c_{g_0 j} \right) \left(2 z_{0j} - c_{g_0 j} - c_{g_i j} \right) \right\}$$

$$= H_1 + \frac{H_2}{\alpha}$$

where,

$$H_1 == \frac{m}{t_{opt}} \sum_{j=1}^{t_{opt}} \left\{ \left(c_{g_i j} - c_{g_0 j} \right) \left(2 z_{0j} - c_{g_0 j} - c_{g_i j} \right) \right\};$$

$$H_2 == \frac{m-1}{2} \sum_{j=1}^{t_{opt}} \left\{ \left(\overline{\beta}_{t_{opt}} - \beta_j \right) \left(c_{g_i j} - c_{g_0 j} \right) \left(2 z_{0j} - c_{g_0 j} - c_{g_i j} \right) \right\};$$

Figure 6.3 shows two standard hyperbolic functions with $H_1 = 0$ and $H_2 = \pm 1$. Hyperbolic function is always monotonic in each branch. We will utilize this monotonic feature to prove 6.22.

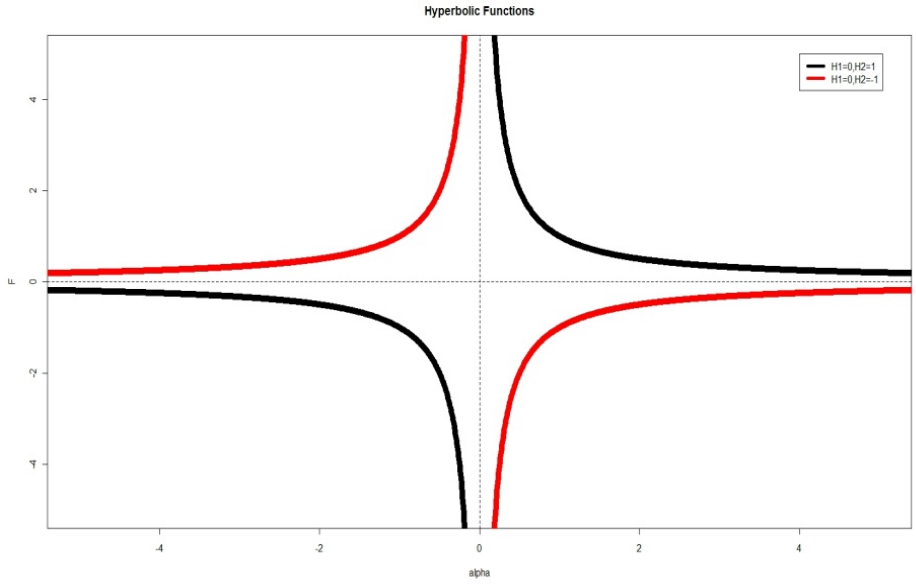


Figure 6.3 Hyperbolic function with H_1=0, H_2=1 (black) and H_1=0, H_2=-1 (red)

We assume the uniqueness of cluster assignment. The proof contains two part with positive and negative scale parameter H2

1. *When $H_2 > 0$, according to monotonic feature of hyperbolic function, $F_{z_0,g_0,g_i}(\alpha)$ is strictly decreasing when α >0. To prove 6.22, we only need to show that:*

$$F_{z_0,g_0,g_i}\big(\alpha_{min} = g(t)\big) < 0 \;\; if \; \exists \alpha^* \in (g(t), g(t+1)], s.t. F_{z_0,g_0,g_i}(\alpha^*) < 0\,.$$

Proof by contradiction: Assuming $\exists\, \mathrm{i} \neq 0, F_{z_0,g_0,g_i}\big(\alpha_{min} = g(t)\big) > 0$, *but* $F_{z_0,g_0,g_i}(\alpha^*) < 0$.*That means* $z_0 = \{z_{0j}\} \notin C_{g_0}$ *when* $\alpha_{min} = g(t)$. *So there must be another cluster partition* $\mathrm{C}' \neq \mathrm{C}$ *existing so that* $F_{z_0,g_0',g_i'}\big(\alpha_{min} = g(t)\big) < 0 \; for \; \forall i \neq 0$. *Then because* $\alpha^* > \alpha_{min} = g(t)$, $F_{z_0,g_0',g_i'}(\alpha^*) < 0 \; for \; \forall i \neq 0$. *On the other hand, we already know* $F_{z_0,g_0,g_i}(\alpha^*) < 0$ *as well, that means* $z_0 \in C_{g_0'}$ *and also* $z_0 \in C_{g_0}$ *at the same time, which contradicts with the uniqueness assumption.*

2. *When $H_2 < 0$, $F_{z_0,g_0,g_i}(\alpha)$ is strictly increasing when α >0. Similarly, we can prove 6.22 is true by showing:*

$$F_{z_0,g_0,g_i}\big(\alpha_{max} = g(t+1)\big) < 0 \;\; if \; \exists \alpha^* \in (g(t), g(t+1)], s.t. F_{z_0,g_0,g_i}(\alpha^*) < 0\,.$$

Now we proved the Proposition 6.2, which implies that, in terms of finding the best cluster assignment, the selection of α is equivalent to determination of optimal number of nonzero variable weights t. since t only have limited possible choices from 1

to m while α could be any positive real number, it is much more efficient to select t in a finite space instead of selecting α in infinite space.

Determining t is similar to the concept of feature selection, but not exactly the same. In feature selection, the aim is to reduce the dimensionality and remove redundant information while here we are only interested in detecting noisy variables which provide little information. Information redundancy is not an issue in our case. For example, if two variables are highly correlated, only one is selected in feature selection to lower the dimension, but we assign nonzero weights for both variables in weighted k-means clustering. Therefore, feature selection techniques are not a good choice to determine *t*. Thus we proposed an efficient measurement called "Reduced Variation" (RV) to determine the number of nonzero variable weights *t*. RV of *ith* variable is defined as follow:

$$RV_i = \frac{1-\beta_i}{\sum_{i=1}^{m}(1-\beta_i)}; \sum_{i=1}^{m} RV_i = 1 \tag{6.23}$$

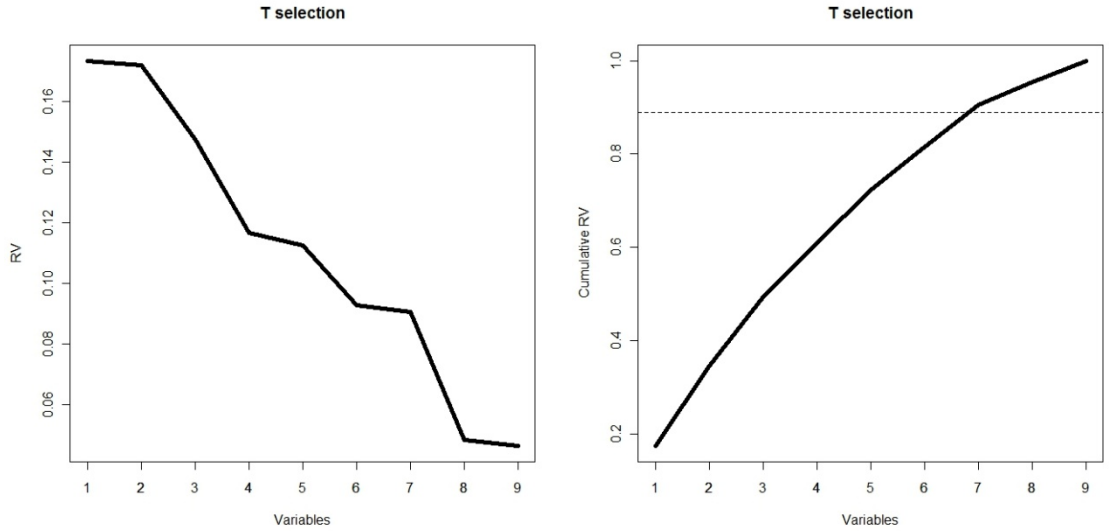


Figure 6.4 An example of Reduced Variation (RV) (left) and cumulative RV (right)

Here β_i is the same as defined before. Since the data is standardized to have unit variance, $1-\beta_i$ is actually the reduced variation on ith variable due to clustering. As shown in Figure 6.4, if a variable is different in different groups, then variation should be reduced a lot by clustering and RV is relatively large; in the other hand, if a variable

remains the same level in all groups, then variation cannot be further reduced by clustering and RV is very small and close to zero.

However, RV is a relative measurement which depends on a lot of things, such as number of variables, number of clusters and signal to noisy ratio in the data. However, there is no unique perfect criterion to determine if RV is large enough. In such variable selection problem, there is always an argument about the balance between removing noise and losing information. From our experience on various datasets, the threshold of $1 - 1/m$ on cumulative RVs always have stable performances in terms of removing the noisy variables without losing too much information. That is, we will select t such that the cumulative RV on first t variables is large than $1 - 1/m$. Then after determining t, we know that all $\alpha \in (g(t), g(t+1)]$ are suitable and, according to proposition 6.2, they will give the same clustering partition which is the best one. So in practice, we will just choose the mean α (6.24).

$$t_{selected} = min\left\{t \middle| \sum_{i=1}^{t} RV_i > \frac{m-1}{m}\right\};$$
$$\alpha_{selected} = \frac{g(t_{selected}) + g(t_{selected}+1)}{2} \tag{6.24}$$

Chapter 7

Results and Comparisons

In this chapter, we will apply our method in two simulated datasets and two real datasets from UCI Machine Learning Repository (Frank and Asuncion 2010) to illustrate the performance of our method. Due to the purpose of comparison, the simulated data 1 and the first real dataset we used are the same datasets Huh and Lim used in their paper. To better illustrate the method, we add one more simulated data and one more real data with higher dimensionality. In section 7.1, four data sets are introduced. Then our method is demonstrated on one data set in detail in section 7.2. Section 7.3 dedicates to the comparison with the original method proposed by Hub and Lim (2009).

7.1 Data Description

Simulated data 1 (Figure 7.1A): this data consists of five 3-dimensional Gaussian groups with 100 observations in each group. Three variables include two informative variables and one noisy variable. Five group means are (5,0,0), (-5,0,0), (0,5,0), (0,-5,0), (0,0,0) with variable-wise standard deviation followed independent normal distribution $N(0,1)$.

Iris data: this is a well-known dataset in pattern recognition literature. The dataset, created by Fisher R.A. in 1936, contains 150 instances of three types of iris, 50 instances each. For each instance, sepal length, sepal width, petal length and petal width were measured in cm as 4 variables.

Simulated data 2 (Figure 7.1B): This data consists of seven 8-dimensional Gaussian groups with 100 observations in each group. Three informative variables and

5 noisy variables are generated. Seven group centers are listed in Table 7.1. Then for each variable, white noise was added which follows standard normal distribution *N(0,1)*.

Table 7.1 Centers of seven groups in simulated data 2

Group 1: (-10, 0, 0, 0, 1, 1, 0, 0)	Group 2: (10, 0, 0, 0, 1, 1, 0, 0)
Group 3: (0, -10, 0, 0, 1, 1, 0, 0)	Group 4: (0, 10, 0, 0, 1, 1, 0, 0)
Group 5: (0, 0, -10, 0, 1, 1, 0, 0)	Group 6: (0, 0, 10, 0, 1, 1, 0, 0)
Group 7: (0, 0, 0, 0, 1, 1, 0, 0)	

Breast tissue data: This data was first published in 1996 (Jossinet 1996). In this data, 4 classes of breast tissue were studied using electrical impedance measurements with multiple frequencies. Then these measurements were transformed into 9 impedance spectrum parameters from where the breast tissue features can be computed. There are 106 observations in total of 4 classes.

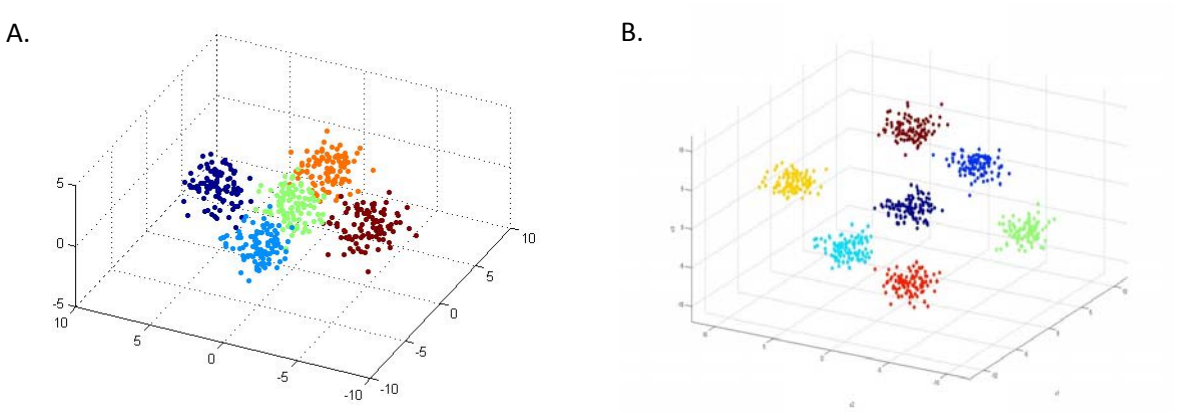


Figure 7.1 Simulated data sets overview: A) simulated data 1 plotted on all three variables; B) simulated data 2 plotted on three informative variables

Four datasets are summarized below (Table 7.2). We will demonstrate the details of our method in the first simulated dataset to show how it works. Then we will compare our method with Huh and Lim's method on all four datasets and illustrate the advantage of our method. We adopt statistical software R (Version 2.10.0) to finish all the work.

Table 7.2 Summary of four datasets

	Classes	Variables	Noise Variables	Observations
Simulation 1	5	3	1	500
Iris	3	4	NA	150
Simulation 2	7	8	5	700
Breast Tissue	4	9	NA	106

7.2 How Our Method Works

Here we first give a detailed example on how our method works using simulated data 1. The data is standardized variable-wise to have zero mean and 1 standard deviation on each variable. After standardization, we first estimate the initial βs as described in section 6.3. Since we only have 3 variables in this case, a {3, 2} simplex lattice with center is formulated to generate seven sets of variable weights. For each set of variable weights, we run the k-means clustering on weighted variable and calculate the overall within cluster sum of squares and then the response variable y in model 6.18. Then after running k-means clustering on all seven sets, we fit the model 6.18 to get the initial estimation on βs and calculate α using 6.24. Finally, we follow iteration algorithm to refine β estimation and calculate the optimal variable weights and clustering partition subsequently.

Table 7.3 β estimation and α selection on simulated data 1

	β	$\mathrm{t}_{\mathrm{selected}}$	$(g(t), g(t+1)]$	$\alpha_{\mathrm{selected}}$
Initial	0.1244,0.1306,0.2313	2	(0.0021, 0.0692]	0.0356
1st iteration	0.0874,0.0827,0.9881	2	(0, 0.6020]	0.3010
2nd iteration	0.0874,0.0827,0.9881	2	(0, 0.6020]	0.3010

For this data, our iteration algorithm takes only two iterations (Table 7.3). Each iteration takes less than 1 second. The *β* of first two informative variables are very small β while the β of third noisy variable is almost 1, which is as expected. Also the algorithm correctly indicates $t_{selected} = 2,$ which is the true number of informative variables. With

the refined estimation on βs and α, we calculated the optimal variable weighting derived in equation 6.16 and compared with the true variable weighting. Since we simulated the data with two informative variables which are different among groups and one additional noisy variable which has nothing to do with the grouping, the true weights should be (1.5, 1.5, 0) for this data. As shown in Table 7.4, our estimated variable weights are almost the same as the true weights.

Table 7.4 Estimated variable weights for simulated data 1

Estimated Variable Weights	$(1.49, 1.51, 0)$
True Variable Weights	$(1.50, 1.50, 0)$

Then we performed the weighted k-means clustering with the estimated variable weights, and the cluster partition is shown in classification table (Table 7.5). We can see five groups are almost separated in five clusters, except 1 member in group 5 is clustered in cluster 3 instead, which results in 99.8% accuracy.

Table 7.5 Cluster partition of weighted k-means clustering for simulated data 1

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Group 1	100	0	0	0	0
Group 2	0	0	0	100	0
Group 3	0	0	100	0	0
Group 4	0	100	0	0	0
Group 5	0	0	1	0	99

We also plotted the weighting curve against a set of α from 2-5 to 25 with increment of 0.01 (Figure 7.2). In the original method proposed by Huh and Lim (2009), the weighting curve was used to determine α. In our method, weighting curve is used for determining α. However, it is good to use weighting curve to examine the calculated weights. Recall in equation 5.9, when $\alpha = 0$, it becomes a linear function on variable weights with all positive coefficient, therefore 5.9 is minimized with one weight equal to m and all others equal to 0; when α increases, the penalty part is emphasized and

therefore all the weights are forced to gradually move towards 1. Such movement is characterized by the way how equation 5.9 is formulated. In Figure 7.2, we can see the featured movement is completed captured, which, in some extent, also confirmed our algorithm.

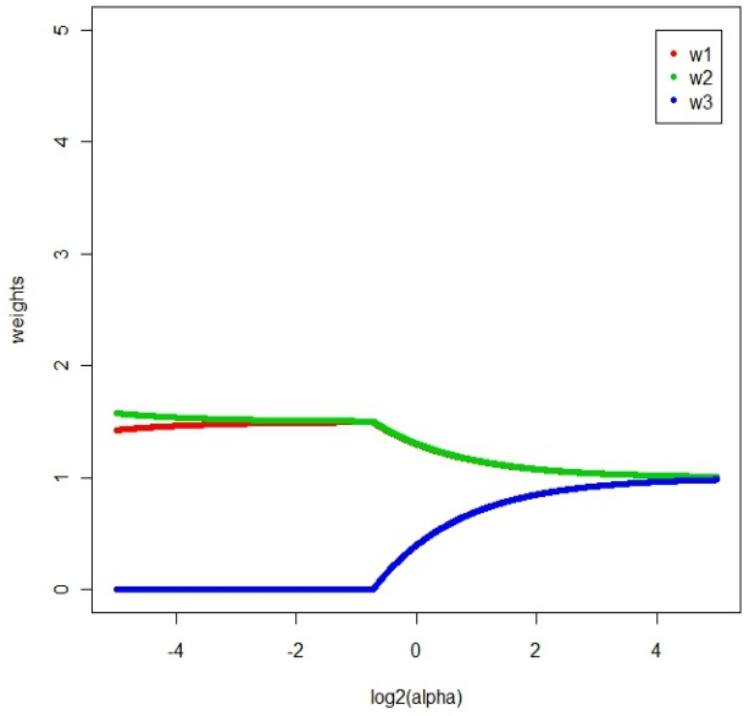


Figure 7.2 Weighting curves of weighted k-means clustering for simulated data 1

The estimated β and selected penalty parameter α obtained using our iteration algorithm for all four data sets are given in Table 7.6. In next section, we will give a detailed comparison between our method and Hub and Lim's method.

7.3 Comparisons to Existing Methods

In this section, we will compare the performance of our method and the method proposed by Huh and Lim (2009). The Nelder-Mead simplex method they used is not global optimal guaranteed and is designed for unconstrained optimization problems only,

which is not our case. We would expect that their method may not be able to find the global optimal solution and subsequently ends up with a suboptimal clustering partition. Therefore here we mainly focus on the algorithm stability and clustering accuracy in the comparison.

Table 7.6 Estimated β and penalty parameter α for four data sets

Data	estimated β	Iteration	α_{selected}
Simulated 1	(0.087,0.083,0.988)	2	0.30
Iris	(0.347, 0.576 0.060, 0.062)	2	0.34
Simulated 2	(0.033, 0.032, 0.031, 0.994, 0.988, 0.992, 0.995, 0.992)	4	0.63
Breast Tissue	(0.064, 0.352, 0.669, 0.273, 0.501, 0.377, 0.185, 0.425, 0.086)	3	1.07

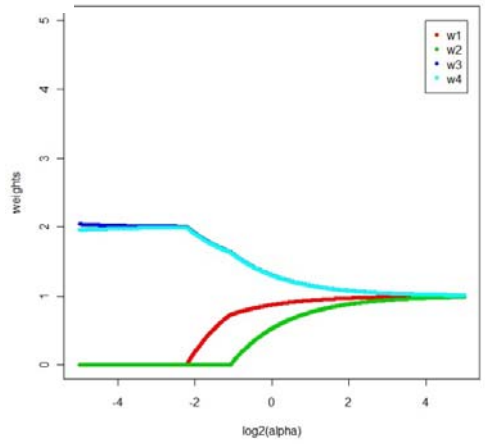
7.3.1 Algorithm Stability

Here we compare the algorithm stability by plotting the weighting curves on a set of penalty parameter α range from 2^{-5} to 2^{5} with increment of 0.01. This is very critical, especially for their method, because in their method, this graph is used to locate a feasible range of α with stable variable weighting. If the algorithm is not stable, then it will be very difficult to even find the feasible range.

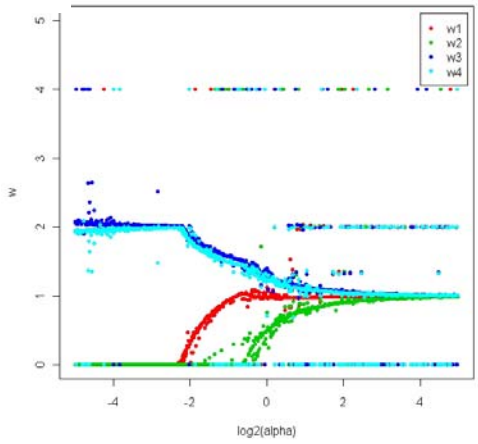
As we discussed in previous section, the objective function determines that all weights will gradually move towards 1 with the increase of penalty parameter α. Figure 7.3 gives the weighting curves for all four datasets generated by our method and Huh and Lim's method. For our method, the weighting curves are generated using the estimated β in Table 7.6 with different penalty parameter α. We can see our method captured this movement very nicely in all four datasets. All the weighting moved slowly towards 1. However, in Huh and Lim's method, this movement was captured only in simulated data 1 and iris data with some outliers. In simulated data 2 and breast tissue data, both with more than 5 variables, their method failed to capture this movement. Instead of expected weighting curve, the plot just looks like random points. The reason is because the Nelder-Mead method is designed only for unconstrained optimization

problems. In our problem, each variable weight is required to be bounded in [*0, m*]. Thus the Nelder- Mead method fails to find the global optimal, and instead, stops at a local optimal solution. For datasets with small number of variables, such as the iris and the simulated data 1, such problem is not severe as their method can still capture the trend. Those outliers can be easily reduced by increasing the increment and reducing the number of points to get a relatively nice-looking and clear curve subsequently. However, when dimensionality increases, such as the simulated data 2 and the breast tissue data, the problem is aggravated as their method can hardly reflect the trend. In this case, one can still reduce the outliers and get a clear plot by reducing the number of points (for example, 8~10 points only). However, it is just not the true weighting curve at all.

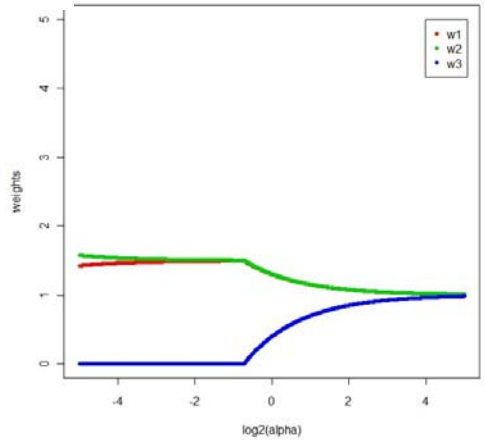
A1.



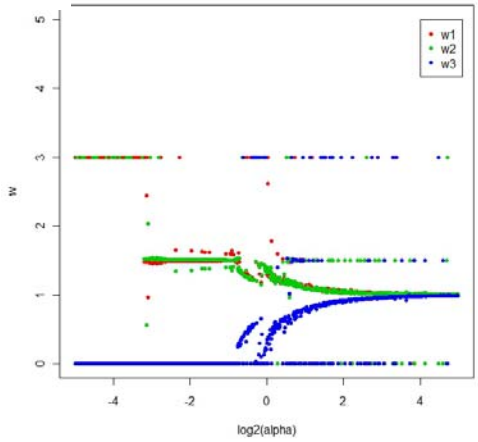
A2.



B1.



B2.



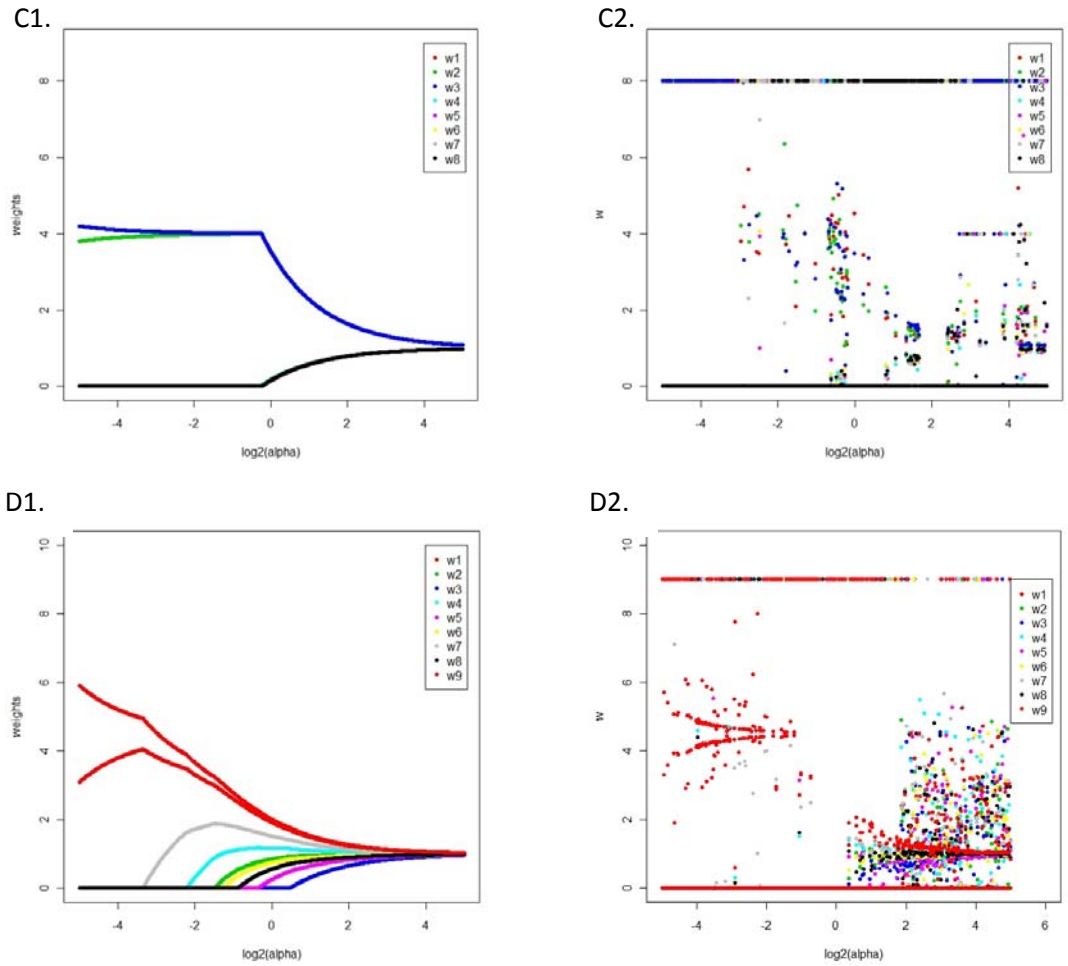


Figure 7.3 Weighting curve comparison between (1) our method and (2) Huh & Lim's method on four datasets: A: Simulated data 1; B: Iris data; C: Simulated data 2; D: Breast tissue data.

7.3.2 Clustering Accuracy

To investigate the clustering accuracy, we calculated the optimal variable weights using both methods and then compared the cluster results, misclassification rate and estimated variable weights of both methods. For our method, the optimal variable weights and corresponding cluster results are generated using the parameters listed in Table 7.6. For Hub and Lim's method, penalty parameter α also need to be determined. However, the graphical method they proposed fails to apply here because of the low quality weighting curve. For comparison purpose, we used the same α in Table 7.6 to perform Hub and Lim's method.

Detailed cluster results are presented at the end of this Chapter (Figure 7.4). Since we know the true class information for all four datasets, here we first compare the clustering performance using misclassification rate (MR). MR is defined as follow: first, each cluster is labeled as the group of majority members in the cluster and members from groups rather than the labeled group are considered misclassified; then we count the misclassified cluster members in all clusters and divide it by the total number of observations as misclassification rate (MR). MRs are compared in Table 7.7. We presented the misclassification rate in both ratio and percentage. We can see that in simulated data 1, both methods did a great job with only 1 (out of 500) misclassified observation. However, in all other three datasets, our method performs better with at least 50% lower in misclassification rate compared to their method.

Table 7.7 Misclassification rate comparison for two methods

Data		Simulated 1	Iris	Simulated 2	Breast Tissue
		m=3,k=5	m=4,k=3	m=8,k=7	m=9,k=4
Our Method	MR	1/500	6/150	0/700	19/106
	MR in %	0.2%	4.0%	0.0%	17.9%
Huh and Lim's Method	MR	1/500	14/150	189/700	35/106
	MR in %	0.2%	9.3%	27.0%	33.0%

The corresponding variable weights are listed below (Table 7.8). In simulated data 1, both methods find the same variable weighting. In all other three datasets, our method find different variable weighting with their method, which results in difference in the clustering partition and misclassification rate. It is interesting to point out that a slightly difference in variable weighting could lead to a big difference in clustering partition as seen in Iris data. Also please note, in both simulated datasets, our method is always able to distinguish the informative variables and noisy variables by assigning different weights. But their method fails for the simulated data 2. Given the way the simulated data sets are generated, the true variable weights should be (1.5, 1.5, 0) for simulated data 1 and (2.67, 2.67, 2.67, 0, 0, 0, 0, 0) for simulated data 2 as the noisy

variables have weight 0 and informative variables have same weight. Therefore, our method almost successfully find the true variable weights for both simulated data while Huh and Lim's method only be able to approach the truth for simulated data 1.

Table 7.8 Optimal variable weighting calculated from two methods

	Our Method	Huh & Lim's Method
Simulated Data 1	(1.49, 1.51, 0)	(1.49, 1.51, 0)
Iris	(0.50, 0, 1.75, 1.75)	(0.58, 0, 1.75, 1.67)
Simulated Data 2	(2.66, 2.67, 2.67, 0, 0, 0, 0, 0)	(0, 3.57, 4.43, 0, 0, 0, 0, 0)
Breast Tissue	(1.94, 0.87, 0, 1.16, 0.32, 0.77, 1.49, 0.59, 1.86)	(1.62, 0.06, 0.01, 1.04, 1.96, 0.67, 1.25, 0.95, 1.54)

The comparisons shown above confirmed that our method is better than Huh and Lim's method in both algorithm stability and clustering accuracy as indicated by weighting curve, misclassification rate and optimal variable weights. This result is somehow expected due to the nature of the optimization method employed. In low dimensional data, Nelder-Mead method performs relatively well. With the increase of dimensionality, Nelder-Mead Simplex method yields poor performance and is unable to find the global optimal in constrained optimization problems. But in our method, we first derive the optimal variable weights theoretically, then we and provide a quantitative method to determine the penalty parameter, which guarantee the performance of our method.

A1.

	C1	C2	C3	C4	C5
T1	100	0	0	0	0
T2	0	0	100	0	0
T3	0	100	0	0	0
T4	0	0	0	0	100
T5	0	1	0	99	0

A2.

	C1	C2	C3	C4	C5
T1	0	0	0	0	100
T2	100	0	0	0	0
T3	0	0	100	0	0
T4	0	0	0	100	0
T5	0	99	1	0	0

B1.

	C1	C2	C3
T1	0	50	0
T2	49	0	1
T3	5	0	45

B2.

	C1	C2	C3
T1	0	50	0
T2	1	0	49
T3	37	0	13

C1.

	C1	C2	C3	C4	C5	C6	C7
T1	0	0	0	100	0	0	0
T2	100	0	0	0	0	0	0
T3	0	0	100	0	0	0	0
T4	0	0	0	0	0	100	0
T5	0	100		0	0	0	0
T6	0	0	0	0	0	0	100
T7	0	0	0	0	100	0	0

C2.

	C1	C2	C3	C4	C5	C6	C7
T1	0	0	0	0	0	51	49
T2	0	0	0	0	0	59	41
T3	0	0	0	0	0	62	38
T4	0	0	0	57	43	0	0
T5	0	100		0	0	0	0
T6	100	0	0	0	0	0	0
T7	0	0	100	0	0	0	0

D1.

	C1	C2	C3	C4
T1	18	3	0	0
T2	2	47	0	0
T3	0	4	0	10
T4	0	0	8	14

D2.

	C1	C2	C3	C4
T1	0	21	0	0
T2	0	49	0	0
T3	0	4	10	0
T4	7	0	14	1

Figure 7.4 Cluster results of both (1) Our method and (2) Huh & Lim's method on four datasets: A: Simulated data 1; B: Iris data; C: Simulated data 2; D: Breast tissue data. T stands for "True group" and C stands for "Cluster"

Chapter 8

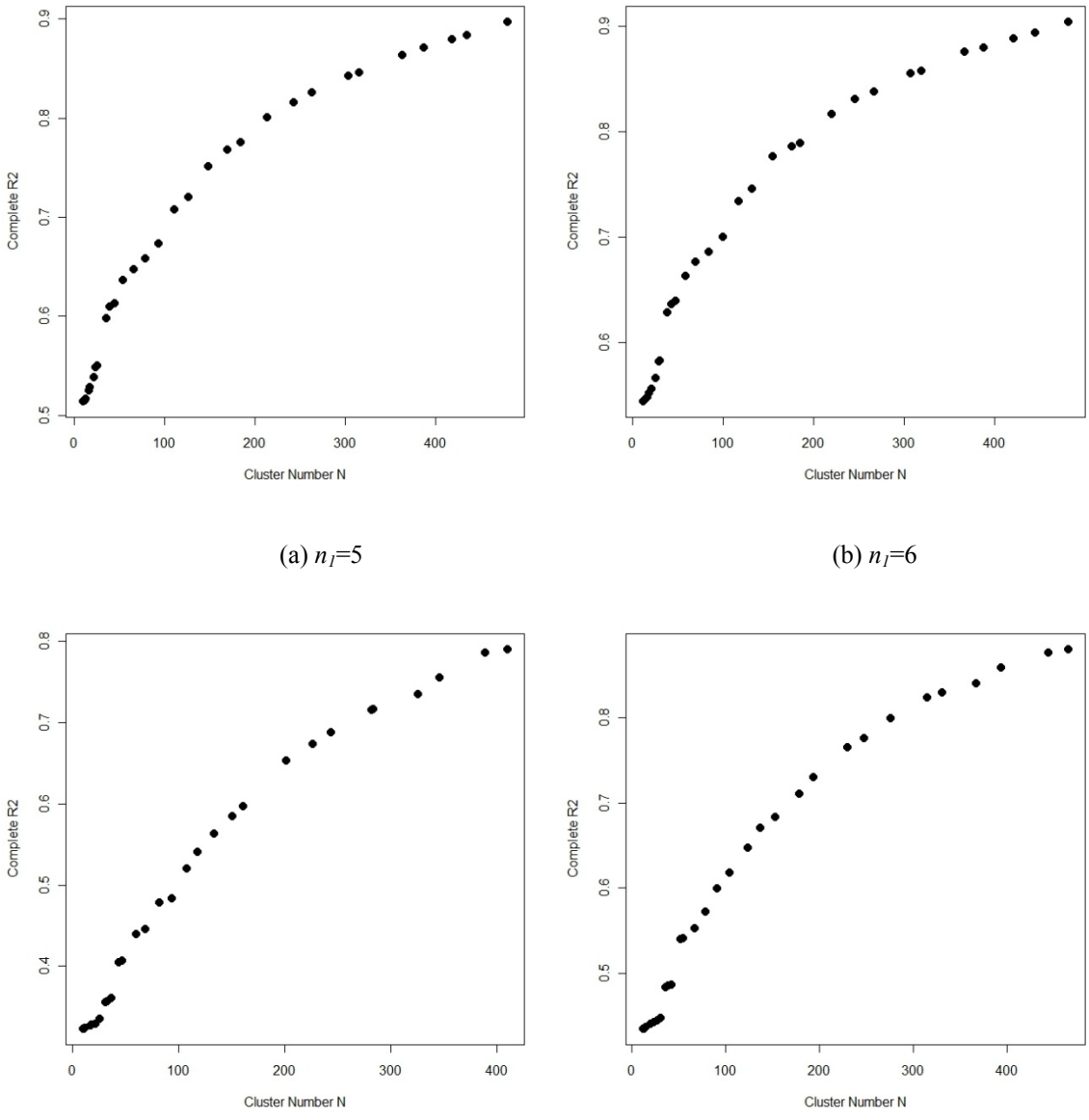
Discussion and Future Work

8.1 Compound Clustering and Constrained Clustering

In part I of this thesis work, we developed two new multi-objective clustering methods: compound clustering and constrained clustering to cluster objects based on multiple objectives using multiple data sources. One real dual-objective application on microarray data was provided to illustrate the methodology. It demonstrated that with appropriate distance measures for both gene expression data and binary biological functional data, the newly proposed compound clustering is superior, compared to traditional hierarchical clustering, to find desired clusters with dual-objective: similar gene expression profile and common biological functions. Furthermore, the newly proposed statistic "complete linkage" R^2 appears to be suitable for cluster number determination in this case. The methodology of compound clustering and the statistic "complete linkage" R^2 for cluster number determination on dual-objective case can be easily extended to a general n-objective problem.

While we have successfully completed the development of compound clustering, there are spaces for the development of constrained clustering. Constrained clustering is a step-by-step approach. A cluster number needs to be specified for each step before moving on to the next step. Given the nature of the nested structure and the lack of an overall distance measurement, It is much harder to define an overall statistic for cluster number determination comparable to the "complete linkage" R^2 for the compound clustering, for the entire constrained clustering procedure. We illustrate our point with a dual-objective constrained clustering analysis on the same temporal gene microarray data from Cold Spring Harbor Laboratory. The "Complete linkage" R^2 is used as a goodness of fit measure. For constrained clustering, we need to determine (n_1, n). Since

this is a nested two-step algorithm, the total cluster number n will exponentially increase with the cluster number n_1 in the first step, therefore n_1 cannot be very large (<10). We found, as expected, that the R^2 is not suitable for this case. As shown in Figure 8.1, by roughly setting n_1=5, 6, we can see that the R^2 is still very low (<0.7) even when the cluster number is over 100.



(a) n_1=5 (b) n_1=6

(c) n_1=5 (d) n_1=6

Figure 8.1 R^2 vs. cluster number under different initial conditions. (a), (b) are for BAS and (c), (d) are for PKJ cells. The x axis is the cluster number, and the y axis is the R^2.

Finding the appropriate way to determine cluster number is the first priority for our future work on constrained cluster analysis. One possible solution is to determine cluster number dynamically and iteratively. Also in step 2, we can determine cluster number for each cluster resulted from step 1 individually. However, both ideas require significant computation time which is not favorable. Once it is done, we can also embark on the comparison of the compound and constrained cluster analysis methods to other multiple-objective clustering analysis methods – using common data sets. Furthermore, we can also extend the application of compound /constrained framework into other fields with multiple data sources, rather than bioinformatics. This work has the potential to lend insight to further development of the multiple-objective clustering framework.

8.2 Weighted K-means Clustering

In part II of this thesis work, we derived the close-form theoretical solution of optimal variable weights for the weighted k-means clustering analysis with a penalized objective function proposed by Huh and Lim (2009). Then we proposed an EM-like iteration algorithm to numerically solve for the optimal variable weights and subsequent, the clustering result. The performance of the proposed method has been demonstrated with two simulated datasets and two real datasets in this paper.

The performance of Huh and Lim's method became poor with increase in data dimensionality. Our proposed method outperformed Huh and Lim's method on both algorithm stability and clustering accuracy, especially with higher dimensional data. It is also interesting to note that our method, by deriving the optimal weights in closed form theoretically, can also be implemented efficiently to estimate the optimal weights. Such time efficiency provides an opportunity for our method to be applied on high dimensional data, such as genetic research and bioinformatics.

To further reduce the time cost when applying to high dimensional data, Principal Component Analysis can be used first to reduce the dimensionality first and then, for a desired cluster number K, our weighted k-means method can be performed on first K-1

principal components instead. Ding and He's paper (Ding and He 2004), which discussed the connection between PCA and k-means clustering, provides the theoretical support for such strategy.

Furthermore, as part of the development, we proposed the cumulative reduced variation (RV) to quantitatively determine the penalty parameter α. Further statistical inference is encouraged to perform on penalty parameter α.

In the examples, we only illustrated our method with correct cluster number K. However, since k-means clustering requires the pre-determined cluster number K, mis-specified cluster number K could have a significant effect on the final cluster result. Therefore, the effect of mis-specified cluster number K should also be investigated for our method.

Finally, the optimal variable weighting we discussed here is based on the objective function which penalize for heterogeneity in variable weights, as shown below. One also can investigate the optimal variable weights from different point of view with other types of penalty functions.

$$\sum_{g=1}^{k}\sum_{i\in I_g}\sum_{j=1}^{m}\frac{w_j(z_{ij}-c_{gj})^2}{n-1}+\alpha\sum_{j=1}^{m}\frac{(w_j-1)^2}{m-1}$$

Bibliography

Acinas, SG, Klepac-Ceraj, V, Hunt, DE, Pharino, C, et al. (2004), "Fine-Scale Phylogenetic Architecture of a Complex Bacterial Community," *Nature*, 430, 551-554.

Agrawal, R, Gehrke, J, Gunopulos, D, and Raghavan, P. (2005), "Automatic Subspace Clustering of High Dimensional Data," *Data Mining and Knowledge Discovery*, 11, 5-33.

Andreopoulos, B, An, A, Wang, X, and Schroeder, M. (2009), "A Roadmap of Clustering Algorithms: Finding a Match for a Biomedical Application," *Brief Bioinform*, 10, 297-314.

Andritsos, P, Tsaparas, P, Miller, R, and Sevcik, K. (2004), "Limbo: Scalable Clustering of Categorical Data

Advances in Database Technology - Edbt 2004," (Vol. 2992), eds. E Bertino, S Christodoulakis, D Plexousakis, V Christophides, M Koubarakis, K Böhm and E Ferrari, Springer Berlin / Heidelberg, pp. 531-532.

Ankerst, M, Breunig, MM, Kriegel, H-P, J\, et al. (1999), "Optics: Ordering Points to Identify the Clustering Structure," *SIGMOD Rec.*, 28, 49-60.

Arabie, P, and Hubert, L. (1994), "Cluster Analysis in Marketing Research," in *Advanced Methods in Marketing Research*, ed. RP Bagozzi, Oxford: Blackwell, pp. 160-189.

Arlia, D, and Coppola, M. (2001), "Experiments in Parallel Clustering with Dbscan," *Proceedings of the 7th International Euro-Par Conference Manchester on Parallel Processing*, 326-331.

Ashburner, M, Ball, CA, Blake, JA, Botstein, D, et al. (2000), "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium," *Nat Genet*, 25, 25-29.

Astrahan, MM. (1970), "Speech Analysis by Clustering, or the Hyperphoneme Method," Technical, Standford University].

Baeza-Yates, RA. (1992), "Introduction to Data Structures and Algorithms Related to Information Retrieval," in *Information Retrieval*, eds. WB Frakes and RA Baeza-Yates, Prentice-Hall, Inc., pp. 13-27.

Bagui, SC. (2005), "Combining Pattern Classifiers: Methods and Algorithms," *Technometrics*, 47, 517-518.

Bandyopadhyay, S, Mukhopadhyay, A, and Maulik, U. (2007), "An Improved Algorithm for Clustering Gene Expression Data," *Bioinformatics*, 23, 2859-2865.

Banerjee, A, Krumpelman, C, Ghosh, J, Basu, S, and Mooney, RJ. (2005), "Model-Based Overlapping Clustering," *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 532-537.

Barker, D. (1976), "Hierarchic and Non-Hierarchic Grouping Methods: An Empirical Comparison of Two Techniques," *Geografiska Annaler. Series B, Human Geography*, 58, 42-58.

Basak, SC, Magnuson, VR, Niemi, GJ, and Regal, RR. (1988), "Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices," *Discrete Applied Mathematics*, 19, 17-44.

Baumgartner, R, Ryner, L, Richter, W, Summers, R, et al. (2000), "Comparison of Two Exploratory Data Analysis Methods for Fmri: Fuzzy Clustering Vs. Principal Component Analysis," *Magnetic Resonance Imaging*, 18, 89-94.

Ben-Hur, A, Horn, D, Siegelmann, HT, and Vapnik, V. (2002), "Support Vector Clustering," *J. Mach. Learn. Res.*, 2, 125-137.

Bernd, F. (1994), "Growing Cell Structures—a Self-Organizing Network for Unsupervised and Supervised Learning," *Neural Networks*, 7, 1441-1460.

Bezdek, JC, Ehrlich, R, and Full, W. (1984), "Fcm: The Fuzzy C-Means Clustering Algorithm," *Computers & Geosciences*, 10, 191-203.

Bezdek, JC, Keller, J, Krisnapuram, R, and Pal, NR (2005), *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing (the Handbooks of Fuzzy Sets)*, Springer-Verlag New York, Inc. .

Bhatia, SK, and Deogun, JS. (1998), "Conceptual Clustering in Information Retrieval," *IEEE Trans. Systems, Man, and Cybernetics*, 28, 10.

Bishop, CM (1995), *Neural Networks for Pattern Recognition*, Oxford Univeristy Press.

Boiman, O, Shechtman, E, and Irani, M. (2008), "In Defense of Nearest-Neighbor Based Image Classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Botev, Z, and Kroese, DP. (2004), "Global Likelihood Optimization Via the Cross-Entropy Method, with an Application to Mixture Models," pp. 517-523.

Boutros, PC, and Okey, AB. (2005), "Unsupervised Pattern Recognition: An Introduction to the Whys and Wherefores of Clustering Microarray Data," *Brief Bioinform*, 6, 331-343.

Bradley, PS, and Fayyad, UM. (1998), "Refining Initial Points for K-Means Clustering," *Proceedings of the Fifteenth International Conference on Machine Learning*, 91-99.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5-32.

Breiman, L, Friedman, JH, Olshen, RA, and Stone, CJ (1984), *Classification and Regression Trees*, Chapman and Hall/CRC.

Brinkmann, WAR. (1999), "Application of Non-Hierarchically Clustered Circulation Components to Surface Weather Conditions: Lake Superior Basin Winter Temperatures," *Theoretical and Applied Climatology*, 63, 41-56.

Brusco, MJ, and Cradit, JD. (2001), "A Variable-Selection Heuristic for K-Means Clustering," *Psychometrika*, 66, 249-270.

Burges, CJC. (1998), "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, 121-167.

Carmone, FJ, Kara, A, and Maxwell, S. (1999), "Hlnov: A New Model to Improve Market Segment Definition by Identifying Noisy Variables," *Journal of Marketing Research*, 36, 501-509.

Carpineto, C, and Romano, G. (1996), "A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval," *Machine Learning*, 24, 95-122.

Carroll, J, and Chaturvedi, A. (1998), "K-Midranges Clustering," in *Advances in Data Science and Classification*, eds. A Rizzi, M Vichi and HH Bock, Berlin: Springer, pp. 3-14.

Caruana, R, and Niculescu-Mizil, A. (2006), "An Empirical Comparison of Supervised Learning Algorithms," *Proceedings of the 23rd international conference on Machine learning*, 161-168.

Chaturvedi, A, Green, PE, and Caroll, JD. (2001), "K-Modes Clustering," *Journal of Classification*, 18, 35-55.

Chen, Y, Zhu, L, Yuille, A, and Zhang, H. (2009), "Unsupervised Learning of Probabilistic Object Models (Poms) for Object Classification, Segmentation, and Recognition Using Knowledge Propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31, 1747-1761.

Cheng, H, Cao, J, Wang, X, and Das, SK. (2006), "Stability-Based Multi-Objective Clustering in Mobile Ad Hoc Networks," *Proceedings of the 3rd international conference on Quality of service in heterogeneous wired/wireless networks*, 27.

Cheng, J, Cline, M, Martin, J, Finkelstein, D, et al. (2004), "A Knowledge-Based Clustering Algorithm Driven by Gene Ontology," *J Biopharm Stat*, 14, 687-700.

Cheng, X, and Wallace, JM. (1993), "Cluster Analysis of the Northern Hemisphere Wintertime 500-Hpa Height Field: Spatial Patterns," *Journal Name: Journal of the Atmospheric Sciences; (United States); Journal Volume: 50:16*, Medium: X; Size: Pages: 2674-2696.

Cheng, Y, and Church, GM. (2000), "Biclustering of Expression Data," *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 93-103.

Chiang, M, and Mirkin, B. (2006), "Determining the Number of Clusters in the Straight K-Means: Experimental Comparison of Eight Options," in *United Kingdom Computational Intelligence Workshop*, Leeds, pp. 143-150.

Chiang, M, and Mirkin, B. (2007), "Experiments for the Number of Clusters in K-Means

Progress in Artificial Intelligence," (Vol. 4874), eds. J Neves, M Santos and J Machado, Springer Berlin / Heidelberg, pp. 395-405.

Chipman, H, and Tibshirani, R. (2006), "Hybrid Hierarchical Clustering with Applications to Microarray Data," *Biostatistics*, 7, 286-301.

Chiu, T, Fang, D, Chen, J, Wang, Y, and Jeris, C. (2001), "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment," *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 263-268.

Clare, A, and King, RD. (2002), "How Well Do We Understand the Clusters Found in Microarray Data?," *In Silico Biol*, 2, 511-522.

Clatworthy, J, Buick, D, Hankins, M, Weinman, J, and Horne, R. (2005), "The Use and Reporting of Cluster Analysis in Health Psychology: A Review," *British Journal of Health Psychology*, 10, 329-358.

Cohen, I, Cozman, FG, Sebe, N, Cirelo, MC, and Huang, TS. (2004), "Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction," *IEEE Trans Pattern Anal Mach Intell*, 26, 1553-1567.

Collobert, R, and Weston, J. (2008), "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," *Proceedings of the 25th international conference on Machine learning*, 160-167.

Cordes, D, Haughton, V, Carew, JD, Arfanakis, K, and Maravilla, K. (2002), "Hierarchical Clustering to Measure Connectivity in Fmri Resting-State Data," *Magnetic Resonance Imaging*, 20, 305-317.

Corne, DW, Jerram, NR, Knowles, JD, and Oates, MJ. (2001), "Pesa-Ii: Region-Based Selection in Evolutionary Multiobjective Optimization," in *Genetic and Evolutionary Computation Conference* pp. 283-290.

Cornell, JA (2002), *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data* (3rd Edition ed.), New York: Wiley.

Cox, DR. (1957), "Note on Grouping," *Journal of the American Statistical Association*, 52, 543-547.

Dahlquist, KD, Salomonis, N, Vranizan, K, Lawlor, SC, and Conklin, BR. (2002), "Genmapp, a New Tool for Viewing and Analyzing Microarray Data on Biological Pathways," *Nat Genet*, 31, 19-20.

Dasarathy, BV (1991), *Nearest Neighbor ({Nn}) Norms: {Nn} Pattern Classification Techniques*, IEEE Computer Society Press.

Datta, S. (2001), "Exploring Relationships in Gene Expressions: A Partial Least Squares Approach," *Gene Expr*, 9, 249-255.

David, LD, and Donald, WB. (1979), "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224-227.

De Backer, S, and Scheunders, P. (1999), "A Competitive Elliptical Clustering Algorithm," *Pattern Recognition Letters*, 20, 1141-1147.

De Soete, G. (1986), "Optimal Variable Weighting for Ultrametric and Additive Tree Clustering," *Quality & Quantity*, 20, 169-180.

De Soete, G, and Carroll, J. (1994), "K-Means Clustering in a Low-Dimensional Euclidean Space," in *New Approaches in Classification and Data Analysis*, eds. E Diday, Y LECHEVALLIER, M Schader, P BERTRAND and B BURTSCHY, Berlin: Springer-Verlag, pp. 212-219.

Deb, K. (2003), "Multi-Objective Evolutionary Algorithms: Introducing Bias among Pareto-Optimal Solutions

Advances in Evolutionary Computing," eds. A Ghosh and S Tsutsui, Springer Berlin Heidelberg, pp. 263-292.

Dembélé, D, and Kastner, P. (2003), "Fuzzy C-Means Method for Clustering Microarray Data," *Bioinformatics*, 19, 973-980.

Dempster, AP, Laird, NM, and Rubin, DB. (1977), "Maximum Likelihood from Incomplete Data Via the Em Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.

DeRisi, JL, Iyer, VR, and Brown, PO. (1997), "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, 278, 680-686.

Desai, D, May, RD, Haldar, P, Shah, S, et al. (2011), "Cytokine Profiling in Severe Asthma Subphenotypes Using Factor and Cluster Analysis," *Am. J. Respir. Crit. Care Med.*, 183, A3719-.

DeSarbo, W, Carroll, J, Clark, L, and Green, P. (1984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables," *Psychometrika*, 49, 57-78.

Desarbo, W, Jedidi, K, Cool, K, and Schendel, D. (1991), "Simultaneous Multidimensional Unfolding and Cluster Analysis: An Investigation of Strategic Groups," *Marketing Letters*, 2, 129-146.

Di Nuovo, AG, Palesi, M, and Catania, V. (2007), "Multi-Objective Evolutionary Fuzzy Clustering for High-Dimensional Problems," in *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pp. 1-6.

Diday, E, and Simon, JC. (1976), "Clustering Analysis," in *Digital Pattern Recognition*, ed. KS Fu, Springer-Verlag, New York, Inc., pp. 47-94.

Dillon, WR, Mulani, N, and Frederick, DG. (1989), "On the Use of Component Scores in the Presence of Group-Structure," *Journal of Consumer Research*, 16, 106-112.

Dimitriadou, E, Dolničar, S, and Weingessel, A. (2002), "An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets," *Psychometrika*, 67, 137-159.

Ding, C, and He, X. (2004), "K-Means Clustering Via Principal Component Analysis," *Proceedings of the twenty-first international conference on Machine learning*, 29.

Divina, F, and Aguilar-Ruiz, JS. (2007), "A Multi-Objective Approach to Discover Biclusters in Microarray Data," *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 385-392.

Dudoit, S, and Fridlyand, J. (2002), "A Prediction-Based Resampling Method for Estimating the Number of Clusters in a Dataset," *Genome Biol*, 3, 1-21.

Dunn, JC. (1974), "Well-Separated Clusters and Optimal Fuzzy Partitions," *Journal of Cybernetics*, 4, 95-104.

Duran, BS, and Odell, PL (1974), *Cluster Analysis: A Survey*, Springer-Verlag New York, Inc.

Dysvik, B, and Jonassen, I. (2001), "J-Express: Exploring Gene Expression Data Using Java," *Bioinformatics*, 17, 369-370.

Eisen, MB, Spellman, PT, Brown, PO, and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc Natl Acad Sci U S A*, 95, 14863-14868.

Engelman, L, and Hartigan, JA. (1969), "Percentage Points of a Test for Clusters," *Journal of the American Statistical Association*, 64, 1647-1648.

Ernst, J, Nau, GJ, and Bar-Joseph, Z. (2005), "Clustering Short Time Series Gene Expression Data," *Bioinformatics*, 21 Suppl 1, i159-168.

Esteban, P, Martin-Vide, J, and Mases, M. (2006), "Daily Atmospheric Circulation Catalogue for Western Europe Using Multivariate Techniques," *International Journal of Climatology*, 26, 1501-1515.

Ester, M, Kriegel, H-p, Jörg, S, and Xu, X. (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 226-231.

Eyke, H. (2005), "Fuzzy Methods in Machine Learning and Data Mining: Status and Prospects," *Fuzzy Sets and Systems*, 156, 387-406.

Faceli, K, de Souto, MCP, de Araújo, DSA, and de Carvalho, ACPLF. (2009), "Multi-Objective Clustering Ensemble for Gene Expression Data Analysis," *Neurocomputing*, 72, 2763-2774.

Figueiredo, MAT, and Jain, AK. (2002), "Unsupervised Learning of Finite Mixture Models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24, 381-396.

Finch, H. (2005), "Comparison of Distance Measures in Cluster Analysis with Dichotomous Data," *Journal of Data Science*, 3, 85-100.

Fisher, RA. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Human Genetics*, 7, 179-188.

Fisher, WD. (1958), "On Grouping for Maximum Homogeneity," *Journal of the American Statistical Association*, 53, 789-798.

Flavián, C, and Polo, Y. (1999), "Strategic Groups Analysis (Sga) as a Tool for Strategic Marketing," *European Journal of Marketing*, 33, 548-569.

Fleury, G, Hero, A, Yoshida, S, Carter, T, et al. (2002), "Clustering Gene Expression Signals from Retinal Microarray Data," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, pp. IV-4024-IV-4027.

Fowlkes, EB, Gnanadesikan, R, and Kettenring, JR. (1988), "Variable Selection in Clustering," *Journal of Classification*, 5, 205-228.

Frank, A, and Asuncion, A. (2010), "Uci Machine Learning Repository ".

Friedman, HP, and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association*, 62, 1159-&.

Friedman, N, and Koller, D. (2003), "Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks," *Machine Learning*, 50, 95-125.

Frigui, H, and Krishnapuram, R. (1999), "A Robust Competitive Clustering Algorithm with Applications in Computer Vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, 21, 450-465.

Garcia-Escudero, LA, and Gordaliza, A. (1999), "Robustness Properties of K Means and Trimmed K Means," *Journal of the American Statistical Association*, 94, 956-969.

Gat-Viks, I, Sharan, R, and Shamir, R. (2003), "Scoring Clustering Solutions by Their Biological Relevance," *Bioinformatics*, 19, 2381-2389.

Gerlinger, C, Wessel, J, Kallischnigg, G, and Endrikat, J. (2009), "Pattern Recognition in Menstrual Bleeding Diaries by Statistical Cluster Analysis," *BMC Women's Health*, 9, 21.

Gibbons, FD, and Roth, FP. (2002), "Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation," *Genome Res*, 12, 1574-1581.

Girman, CJ. (1994), "*Cluster Analysis and Classification Tree Methodology as an Aid to Improve Understanding of Benign Prostatic Hyperplasia*," University of North Carolina at Chapel Hill, Dept. of, Biostatistics.

Gnanadesikan, R, Kettenring, JR, and Tsao, SL. (1995), "Weighting and Selection of Variables for Cluster Analysis," *Journal of Classification*, 12, 113-136.

Gollub, J, and Sherlock, G. (2006), "Clustering Microarray Data," *Methods Enzymol*, 411, 194-213.

GONG, X, and RICHMAN, MB. (1995), "On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies," *Journal of Climate*, 8, 897-931.

Goutte, C, Toft, P, Rostrup, E, Nielsen, FÅ, and Hansen, LK. (1999), "On Clustering Fmri Time Series," *NeuroImage*, 9, 298-310.

Green, PE, and Carmone, FJ. (1969), "Multidimensional Scaling: An Introduction and Comparison of Nonmetric Unfolding Techniques," *Journal of Marketing Research*, 6, 330-341.

Grossman, RL, and Poor, HV. (1996), "Optimization Driven Data Mining and Credit Scoring," in *Computational Intelligence for Financial Engineering, 1996., Proceedings of the IEEE/IAFE 1996 Conference on*, pp. 104-110.

Guha, S, Rastogi, R, and Shim, K. (2000), "Rock: A Robust Clustering Algorithm for Categorical Attributes," *Information Systems*, 25, 345-366.

Guo, H, Renaut, R, Chen, K, and Reiman, E. (2003), "Clustering Huge Data Sets for Parametric Pet Imaging," *Biosystems*, 71, 81-92.

Hajnal, I, and Loosveldt, G. (2000), "The Effects of Initial Values and the Covariance Structure on the Recovery of Some Clsutering Methods," in *Data Analysis, Classification, and Related Methods*, eds. H Kiers, J Rasson, P Groenen and M Schader, Berlin: Springer, pp. 47-52.

Hamers, L, Hemeryck, Y, Herweyers, G, Janssen, M, et al. (1989), "Similarity Measures in Scientometric Research: The Jaccard Index Versus Salton's Cosine Formula," *Information Processing & Management*, 25, 315-318.

Han, J, Kamber, M, and Pei, J (2006), *Data Mining: Concepts and Techniques* (2nd Edition ed.), Morgan Kaufmann.

Handl, J, and Knowles, J. (2004), "Evolutionary Multiobjective ClusteringParallel Problem Solving from Nature - Ppsn Viii," (Vol. 3242), eds. X Yao, E Burke, J Lozano, J Smith, J Merelo-Guervós, J Bullinaria, J Rowe, P Tino, A Kabán and H-P Schwefel, Springer Berlin / Heidelberg, pp. 1081-1091.

Handl, J, and Knowles, J. (2005), "Exploiting the Trade-Off — the Benefits of Multiple Objectives in Data Clustering

Evolutionary Multi-Criterion Optimization," (Vol. 3410), eds. C Coello Coello, A Hernández Aguirre and E Zitzler, Springer Berlin / Heidelberg, pp. 547-560.

Handl, J, and Knowles, J. (2007), "An Evolutionary Approach to Multiobjective Clustering," *IEEE Transactions on Evolutionary Computation*, 11, 56-76.

Hanisch, D, Zien, A, Zimmer, R, and Lengauer, T. (2002), "Co-Clustering of Biological Networks and Gene Expression Data," *Bioinformatics*, 18 Suppl 1, S145-154.

Hastie, T, Tibshirani, R, and Friedman, JH (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition ed.), Springer.

He, Z, Xu, X, and Deng, S. (2002), "Squeezer: An Efficient Algorithm for Clustering Categorical Data," *Journal of Computer Science and Technology*, 17, 611-624.

Heiser, W, and Groenen, P. (1997), "Cluster Differences Scaling with a within-Clusters Loss Component and a Fuzzy Successive Approximation Strategy to Avoid Local Minima," *Psychometrika*, 62, 63-83.

Hinneburg, A, and Keim, D. (1998), "An Efficient Approach to Clustering in Large Multimedia Data Sets with Noise," in *4th International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp. 58-65.

Hormozdiari, F, Alkan, C, Eichler, EE, and Sahinalp, SC. (2009), "Combinatorial Algorithms for Structural Variation Detection in High Throughput Sequenced Genomes," *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*, 218-219.

Hosseini, SMS, Maleki, A, and Gholamian, MR. (2010), "Cluster Analysis Using Data Mining Approach to Develop Crm Methodology to Assess the Customer Loyalty," *Expert Systems with Applications*, 37, 5259-5264.

Hsu, C-C, Chen, C-L, and Su, Y-W. (2007), "Hierarchical Clustering of Mixed Data Based on Distance Hierarchy," *Information Sciences*, 177, 4474-4492.

Huang, C-L, Chen, M-C, and Wang, C-J. (2007a), "Credit Scoring with a Data Mining Approach Based on Support Vector Machines," *Expert Systems with Applications*, 33, 847-856.

Huang, D, and Pan, W. (2006), "Incorporating Biological Knowledge into Distance-Based Clustering Analysis of Microarray Gene Expression Data," *Bioinformatics*, 22, 1259-1268.

Huang, DW, Sherman, B, Tan, Q, Collins, J, et al. (2007b), "The David Gene Functional Classification Tool: A Novel Biological Module-Centric Algorithm to Functionally Analyze Large Gene Lists," *Genome Biol*, 8, 1-16.

Huang, JZ, Ng, MK, Hongqiang, R, and Zichen, L. (2005), "Automated Variable Weighting in K-Means Type Clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27, 657-668.

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

Huang, Z, and Ng, MK. (2003), "A Note on K-Modes Clustering," *Journal of Classification*, 20, 257-261.

Huber, P (1981), *Robust Statistics*, Wiley-Interscience.

Hubert, L, and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193-218.

Huberty, CJ, DiStefano, C, and Kamphaus, RW. (1997), "Behavioral Clustering of School Children," *Multivariate Behavioral Research*, 32, 105-134.

Huh, M-H, and Lim, YB. (2009), "Weighting Variables in K-Means Clustering," *Journal of Applied Statistics*, 36, 67-78.

Huth, R, Nemesova, I, and Klimperová, N. (1993), "Weather Categorization Based on the Average Linkage Clustering Technique: An Application to European Mid-Latitudes," *International Journal of Climatology*, 13, 817-835.

Huttenlocher, DP, Klanderman, GA, and Rucklidge, WA. (1993), "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 850-863.

Ichino, M, and Yaguchi, H. (1994), "Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis," *Systems, Man and Cybernetics, IEEE Transactions on*, 24, 698-708.

Igor, K. (2001), "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective," *Artificial Intelligence in Medicine*, 23, 89-109.

Izmirlian, G. (2004), "Application of the Random Forest Classification Algorithm to a Seldi-Tof Proteomics Study in the Setting of a Cancer Prevention Trial," *Applications of Bioinformatics in Cancer Detection*, 1020, 154-174.

Jain, AK, and Dubes, R (1988), *Algorithms for Clustering Data*, Prentice-Hall, Inc.

Jain, AK, Duin, RPW, and Mao, J. (2000), "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, 22, 4-37.

Jain, AK, Murty, MN, and Flynn, PJ. (1999), "Data Clustering: A Review," *ACM Comput. Surv.*, 31, 264-323.

Jensen, FV (1996), *Introduction to Bayesian Networks*, Springer.

Johnson, SC. (1967), "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241-254.

Jossinet, J. (1996), "Variability of Impedivity in Normal and Pathological Breast Tissue," *Med Biol Eng Comput*, 34, 346-350.

Judd, D, McKinley, PK, and Jain, AK. (1998), "Large-Scale Parallel Data Clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, 20, 871-876.

KALKSTEIN, LS, TAN, G, and SKINDLOV, JA. (1987), "An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification," *Journal of Applied Meteorology*, 26, 717-730.

Kamasak, ME, and Bayraktar, B. (2007), "Clustering Dynamic Pet Images on the Projection Domain," *Nuclear Science, IEEE Transactions on*, 54, 496-503.

Kanehisa, M, Goto, S, Sato, Y, Furumichi, M, and Tanabe, M. (2011), "Kegg for Integration and Interpretation of Large-Scale Molecular Data Sets," *Nucleic Acids Res*.

Kanungo, T, Mount, DM, Netanyahu, NS, Piatko, CD, et al. (2002), "An Efficient K-Means Clustering Algorithm: Analysis and Implementation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24, 881-892.

Karush, W. (1939), "*Minima of Functions of Several Variables with Inequalities as Side Constraints*," University of Chicago, Department of Mathematics.

Kass, GV. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29, 119-127.

Kaufman, L, and Rousseeuw, P (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience.

Kenward, RE, Clarke, RT, Hodder, KH, and Walls, SS. (2001), "Density and Linkage Estimators of Home Range: Nearest-Neighbor Clustering Defines Multinuclear Cores," *Ecology*, 82, 1905-1920.

Kerr, MK, and Churchill, GA. (2001), "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," *Proc Natl Acad Sci U S A*, 98, 8961-8965.

Kerr, MK, Martin, M, and Churchill, GA. (2000), "Analysis of Variance for Gene Expression Microarray Data," *J Comput Biol*, 7, 819-837.

Kiel, GC, and Layton, RA. (1981), "Dimensions of Consumer Information Seeking Behavior," *Journal of Marketing Research*, 18, 233-239.

Kim, D-W, Lee, KH, and Lee, D. (2004), "Fuzzy Clustering of Categorical Data Using Fuzzy Centroids," *Pattern Recognition Letters*, 25, 1263-1271.

Kimura, Y, Hsu, H, Toyama, H, Senda, M, and Alpert, NM. (1999), "Improved Signal-to-Noise Ratio in Parametric Images by Cluster Analysis," *NeuroImage*, 9, 554-561.

King, B. (1967), "Step-Wise Clustering Procedures," *Journal of the American Statistical Association*, 62, 86-101.

Kohonen, T. (1990), "The Self-Organizing Map," *Proceedings of the IEEE*, 78, 1464-1480.

Korkmaz, EE, Du, J, Alhajj, R, and Barker, K. (2006), "Combining Advantages of New Chromosome Representation Scheme and Multi-Objective Genetic Algorithms for Better Clustering," *Intell. Data Anal.*, 10, 163-182.

Kotsiantis, SB. (2007), "Supervised Machine Learning: A Review of Classification Techniques," *Proceeding of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, 3-24.

Kriegel, H-P, Kröger, P, and Zimek, A. (2009), "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," *ACM Trans. Knowl. Discov. Data*, 3, 1-58.

Kuhn, HW, and Tucker, AW. (1951), "Nonlinear Programming," in *2nd Berkeley Symposium on Mathematical Staitistics and Probability*, Berkeley: University of California Press, pp. 481-492.

Larrañaga, P, Calvo, B, Santana, R, Bielza, C, et al. (2006), "Machine Learning in Bioinformatics," *Briefings in Bioinformatics*, 7, 86-112.

Li, C, and Biswas, G. (2002), "Unsupervised Learning with Mixed Numeric and Nominal Data," *Knowledge and Data Engineering, IEEE Transactions on*, 14, 673-690.

Lin, Y, Nadler, S, Lan, H, Attie, A, and Yandell, B. (2003), "Adaptive Gene Picking with Microarray Data: Detecting Important Low Abundance Signalsthe Analysis of Gene Expression Data," eds. G Parmigiani, E Garrett, R Irizarry and S Zeger, Springer London, pp. 291-312.

Liptrot, M, Adams, KH, Martiny, L, Pinborg, LH, et al. (2004), "Cluster Analysis in Kinetic Modelling of the Brain: A Noninvasive Alternative to Arterial Sampling," *NeuroImage*, 21, 483-493.

Lloyd, S. (1982), "Least Squares Quantization in Pcm," *Information Theory, IEEE Transactions on*, 28, 129-137.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symp. Mathematical Statist. Probability*, pp. 281-297.

Mahajan, M, Nimbhorkar, P, and Varadarajan, K. (2009), "The Planar K-Means Problem Is Np-Hard

Walcom: Algorithms and Computation," (Vol. 5431), eds. S Das and R Uehara, Springer Berlin / Heidelberg, pp. 274-285.

Makarenkov, V, and Legendre, P. (2001), "Optimal Variable Weighting for Ultrametric and Additive Trees and ≪I≫K -Means Partitioning: Methods and Software," *Journal of Classification*, 18, 245-271.

Manning, C, Raghavan, P, and Schütze, H (2008), *Introduction to Information Retrieval*, Cambridge University Press.

Matake, N, Hiroyasu, T, Miki, M, and Senda, T. (2007), "Multiobjective Clustering with Automatic K-Determination for Large-Scale Data," *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 861-868.

McLachlan, GJ (2005), *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Inc.

McLachlan, GJ, and Peel, D (2000), *Finite Mixture Models* (1st ed.), Wiley-Interscience.

McShane, LM, Radmacher, MD, Freidlin, B, Yu, R, et al. (2002), "Methods for Assessing Reproducibility of Clustering Patterns Observed in Analyses of Microarray Data," *Bioinformatics*, 18, 1462-1469.

Menard, SW (2001), *Applied Logistic Regression Analysis* (2nd Edition ed.), SAGE Publications.

Messai, N, Devignes, M-D, Napoli, A, and Smail-Tabbone, M. (2008), "Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval," *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, 127-131.

Michael, D. (2000), "The Growing Hierarchical Self-Organizing Map," pp. 6015-6015.

Michelangeli, P-A, Vautard, R, and Legras, B. (1995), "Weather Regimes: Recurrence and Quasi Stationarity," *Journal of Atmospheric Sciences*, 52, 1237-1256.

Milligan, G. (1980), "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika*, 45, 325-342.

Milligan, G, and Cooper, M. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50, 159-179.

Milligan, GW, and Cooper, MC. (1988), "A Study of Standardization of Variables in Cluster Analysis," *Journal of Classification*, 5, 181-204.

Mitchell, TM. (1999), "Machine Learning and Data Mining," *Commun. ACM*, 42, 30-36.

Mitra, S, and Banka, H. (2006), "Multi-Objective Evolutionary Biclustering of Gene Expression Data," *Pattern Recognition*, 39, 2464-2477.

Mo, K, and Ghil, M. (1988), "Cluster Analysis of Multiple Planetary Flow Regimes," *J. Geophys. Res.*, 93, 10927-10952.

Modha, DS, and Spangler, WS. (2003), "Feature Weighting in ≪I≫K -Means Clustering," *Machine Learning*, 52, 217-237.

Mote, TL. (1998), "Mid-Tropospheric Circulation and Surface Melt on the Greenland Ice Sheet. Part Ii: Synoptic Climatology," *International Journal of Climatology*, 18, 131-145.

Nadler, ST, Stoehr, JP, Schueler, KL, Tanimoto, G, et al. (2000), "The Expression of Adipogenic Genes Is Decreased in Obesity and Diabetes Mellitus," *Proc Natl Acad Sci U S A*, 97, 11371-11376.

Nagy, G. (1968), "State-of-the-Art in Pattern Recognition," *Journal Name: Proc. IEEE (Inst. Elec. Electron. Eng.), 56: 836-62(May 1968).; Other Information: Orig. Receipt Date: 31-DEC-68*, Medium: X.

Nakatani, C, and Hirschberg, J. (1993), "A Speech-First Model for Repair Detection and Correction," *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 46-53.

Nelder, JA, and Mead, R. (1965), "A Simplex-Method for Function Minimization," *Computer Journal*, 7, 308-313.

O'Sullivan, F. (1993), "Imaging Radiotracer Model Parameters in Pet: A Mixture Analysis Approach," *Medical Imaging, IEEE Transactions on*, 12, 399-412.

Pan, W. (2005), "Incorporating Biological Information as a Prior in an Empirical Bayes Approach to Analyzing Microarray Data," *Stat Appl Genet Mol Biol*, 4, Article12.

Park, H-S, and Jun, C-H. (2009), "A Simple and Fast Algorithm for K-Medoids Clustering," *Expert Systems with Applications*, 36, 3336-3341.

Peel, D, and McLachlan, GJ. (2000), "Robust Mixture Modelling Using the T Distribution," *Statistics and Computing*, 10, 339-348.

Per-Erik, D. (1980), "Euclidean Distance Mapping," *Computer Graphics and Image Processing*, 14, 227-248.

Pham, DL, Xu, C, and Prince, JL. (2000), "Current Methods in Medical Image Segmentation1," *Annual Review of Biomedical Engineering*, 2, 315-337.

Press, JS, and Wilson, S. (1978), "Choosing between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, 73, 699-705.

Punj, G, and Stewart, DW. (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 20, 134-148.

Quinlan, JR. (1986), "Induction of Decision Trees," *Machine Learning*, 1, 81-106.

Rahman, MM, Bhattacharya, P, and Desai, BC. (2007), "A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques with Relevance Feedback," *Information Technology in Biomedicine, IEEE Transactions on*, 11, 58-69.

Ramette, A. (2007), "Multivariate Analyses in Microbial Ecology," *FEMS Microbiol Ecol*, 62, 142-160.

Rand, WM. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846-850.

Ratnaparkhi, A. (1999), "Learning to Parse Natural Language with Maximum Entropy Models," *Machine Learning*, 34, 151-175.

Redner, RA, and Walker, HF. (1984), "Mixture Densities, Maximum Likelihood and the Em Algorithm," *SIAM Review*, 26, 195-239.

Reidenbach, RE, and Robin, DP. (1990), "Toward the Development of a Multidimensional Scale for Improving Evaluations of Business Ethics," *Journal of Business Ethics*, 9, 639-653.

Ripley, BD. (1994), "Neural Networks and Related Methods for Classification," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 409-456.

Ripley, BD (1996), *Pattern Recognition and Neural Networks*, ed. C Press, Cambridge University Press.

Robinette, SL, Veselkov, KA, Bohus, E, Coen, M, et al. (2009), "Cluster Analysis Statistical Spectroscopy Using Nuclear Magnetic Resonance Generated Metabolic Data Sets from Perturbed Biological Systems," *Analytical Chemistry*, 81, 6581-6589.

Rubinstein, RY, and Kroese, DP (2004), *The Cross Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation (Information Science and Statistics)*, Springer-Verlag New York, Inc.

Ruspini, EH. (1969), "A New Approach to Clustering," *Information and Control*, 15, 22-32.

Salvador, S, and Chan, P. (2004), "Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms," *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, 576-584.

Santos, JA, Corte-Real, J, and Leite, SM. (2005), "Weather Regimes and Their Connection to the Winter Rainfall in Portugal," *International Journal of Climatology*, 25, 33-50.

SAS. (2004), "The Fastclus Procedure," in *Sas/Stat 9.1 User's Guide* (Vol. 2), Cary, NC: SAS Institute.

Saunders, JA. (1980), "Cluster Analysis for Market Segmentation," *European Journal of Marketing*, 14, 422-435.

Schölkopf, B, Smola, A, and Müller, K-R. (1997), "Kernel Principal Component Analysis

Artificial Neural Networks — Icann'97," (Vol. 1327), eds. W Gerstner, A Germond, M Hasler and J-D Nicoud, Springer Berlin / Heidelberg, pp. 583-588.

Serrano, A, García, J, Mateos, VL, Cancillo, ML, and Garrido, J. (1999), "Monthly Modes of Variation of Precipitation over the Iberian Peninsula," *Journal of Climate*, 12, 2894-2919.

Shen, Y, Sun, W, and Li, KC. (2010), "Dynamically Weighted Clustering with Noise Set," *Bioinformatics*, 26, 341-347.

Solman, SA, and Menéndez, CG. (2003), "Weather Regimes in the South American Sector and Neighbouring Oceans During Winter," *Climate Dynamics*, 21, 91-104.

Späth, H (1985), *The Cluster Dissection and Analysis Theory Fortran Programs Examples*, Prentice-Hall, Inc.

Spiller, P, and Lohse, GL. (1997), "A Classification of Internet Retail Stores," *Int. J. Electron. Commerce*, 2, 29-56.

Steinhaus, H. (1957), "Sur La Division Des Corps Matériels En Parties (in French)," *Bull. Acad. Polon. Sci.*, 4, 4.

Steinley, D. (2003), "Local Optima in K-Means Clustering: What You Don't Know May Hurt You," *Psychological Methods*, 8, 294-304.

Steinley, D. (2004), "Standardizing Variables in K -Means Clustering

Classification, Clustering, and Data Mining Applications," (Vol. 0), eds. D Banks, L House, FR McMorris, P Arabie and W Gaul, Springer Berlin Heidelberg, pp. 53-60.

Steinley, D. (2006), "K-Means Clustering: A Half-Century Synthesis," *British Journal of Mathematical and Statistical Psychology*, 59, 1-34.

Steinley, D, and Brusco, M. (2008), "Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures," *Psychometrika*, 73, 125-144.

Strehl, A, and Ghosh, J. (2003), "Cluster Ensembles --- a Knowledge Reuse Framework for Combining Multiple Partitions," *J. Mach. Learn. Res.*, 3, 583-617.

Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, et al. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proc Natl Acad Sci U S A*, 102, 15545-15550.

Suykens, JAK, and Vandewalle, J. (1999), "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, 9, 293-300.

Tari, L, Baral, C, and Kim, S. (2009), "Fuzzy C-Means Clustering with Prior Biological Knowledge," *J Biomed Inform*, 42, 74-81.

Thorndike, R. (1953), "Who Belongs in the Family?," *Psychometrika*, 18, 267-276.

Tibshirani, R, Walther, G, and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set Via the Gap Statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411-423.

Townsend, J. (1971), "Theoretical Analysis of an Alphabetic Confusion Matrix," *Attention, Perception, & Psychophysics*, 9, 40-50.

Tryon, RC (1939), *Cluster Analysis; Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Ann Arbor, Mich.: Edwards brother, inc., lithoprinters and publishers.

Tsai, C-Y, and Chiu, C-C. (2008), "Developing a Feature Weight Self-Adjustment Mechanism for a K-Means Clustering Algorithm," *Computational Statistics & Data Analysis*, 52, 4658-4672.

Tsao, EC-K, Bezdek, JC, and Pal, NR. (1994), "Fuzzy Kohonen Clustering Networks," *Pattern Recognition*, 27, 757-764.

Tseng, GC. (2007), "Penalized and Weighted K-Means for Clustering with Scattered Objects and Prior Information in High-Throughput Biological Data," *Bioinformatics*, 23, 2247-2255.

Tung, F, Wong, A, and Clausi, DA. (2010), "Enabling Scalable Spectral Clustering for Image Segmentation," *Pattern Recognition*, 43, 4069-4076.

Van Buuren, S, and Heiser, W. (1989), "Clustering N Objects into K Groups under Optimal Scaling of Variables," *Psychometrika*, 54, 699-706.

Varela, C, Schmidt, SA, Borneman, AR, Krömer, JO, et al. (2011), "Systems Biology: A New Paradigm for Industrial Yeast Strain Development," *Microbiology Australia*, 32, 151-155.

Verma, B, and Blumenstein, M (2008), *Pattern Recognition Technologies and Applications: Recent Advances* (1st ed.), IGI Global.

Vesanto, J. (2001), "Importance of Individual Variables in the ≪I≫K -Means Algorithm

Advances in Knowledge Discovery and Data Mining," (Vol. 2035), eds. D Cheung, G Williams and Q Li, Springer Berlin / Heidelberg, pp. 513-518.

Vesanto, J, and Alhoniemi, E. (2000), "Clustering of the Self-Organizing Map," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 11, 586-600.

Vichi, M, and Kiers, HAL. (2001), "Factorial K-Means Analysis for Two-Way Data," *Computational Statistics & Data Analysis*, 37, 49-64.

Vlahou, A, Schorge, JO, Gregory, BW, and Coleman, RL. (2003), "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data," *Journal of Biomedicine and Biotechnology*, 308-314.

Vrac, M, Hayhoe, K, and Stein, M. (2007), "Identification and Intermodel Comparison of Seasonal Circulation Patterns over North America," *International Journal of Climatology*, 27, 603-620.

Waddell, PJ, and Kishino, H. (2000), "Cluster Inference Methods and Graphical Models Evaluated on Nci60 Microarray Gene Expression Data," *Genome Inform Ser Workshop Genome Inform*, 11, 129-140.

Ward, JH. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236-244.

Wettschereck, D, Aha, DW, and Mohri, T. (1997), "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms," *Artificial Intelligence Review*, 11, 273-314.

Whitaker, RJ, Grogan, DW, and Taylor, JW. (2003), "Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea," *Science*, 301, 976-978.

Wilson, DR, and Martinez, TR. (1997), "Improved Heterogeneous Distance Functions," *J. Artif. Int. Res.*, 6, 1-34.

Wind, Y. (1978), "Issues and Advances in Segmentation Research," *Journal of Marketing Research*, 15, 317-337.

Wold, S, Esbensen, K, and Geladi, P. (1987), "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, 2, 37-52.

Wong, K-P, Feng, D, Meikle, SR, and Fulham, MJ. (2002), "Segmentation of Dynamic Pet Images Using Cluster Analysis," *Nuclear Science, IEEE Transactions on*, 49, 200-207.

Xing, EP, Ng, AY, Jordan, MI, and Stuart, R. (2003), "Distance Metric Learning, with Application to Clustering with Side-Information," *Learning*, 15, 505-512.

Yeung, KY, Haynor, DR, and Ruzzo, WL. (2001), "Validating Clustering for Gene Expression Data," *Bioinformatics*, 17, 309-318.

Yianilos, PN. (1993), "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces," *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, 311-321.

Yiou, P, and Nogaj, M. (2004), "Extreme Climatic Events and Weather Regimes over the North Atlantic: When and Where?," *Geophys. Res. Lett.*, 31, L07202.

Zeeberg, BR, Feng, W, Wang, G, Wang, MD, et al. (2003), "Gominer: A Resource for Biological Interpretation of Genomic and Proteomic Data," *Genome Biol*, 4, R28.

Zhang, J, and Maringer, D. (2010), "A Clustering Application in Portfolio Management

Electronic Engineering and Computing Technology," (Vol. 60), eds. S-I Ao and L Gelman, Springer Netherlands, pp. 309-321.

Zhang, T-j, Huang, X-h, Tang, J-f, and Luo, X-g. (2011), "Case Study on Cluster Analysis of the Telecom Customers Based on Consumers' Behavior," in *Industrial Engineering and Engineering Management (IE&EM), 2011 IEEE 18Th International Conference on*, pp. 1358-1362.

Zhang, T, Ramakrishnan, R, and Livny, M. (1996), "Birch: An Efficient Data Clustering Method for Very Large Databases," *SIGMOD Rec.*, 25, 103-114.

Zhang, T, Zhu, W, and McGraw, R. (2008), "Joint Cluster and Non-Negative Least Squares Analysis for Aerosol Mass Spectrum Data " *Journal of Physics: Conference Series*, 125.

Zhu, J, and Hastie, T. (2004), "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics*, 5, 427-443.

Zou, M, and Conzen, SD. (2005), "A New Dynamic Bayesian Network (Dbn) Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data," *Bioinformatics*, 21, 71-79.