Chapter 6
Soft Subspace Clustering for Cancer Microarray Data Analysis: A Survey

Natthakan Iam-On
Mae Fah Luang University, Thailand

Tossapon Boongoen
Royal Thai Air Force Academy, Thailand

ABSTRACT

A need has long been identified for a more effective methodology to understand, prevent, and cure cancer. Microarray technology provides a basis of achieving this goal, with cluster analysis of gene expression data leading to the discrimination of patients, identification of possible tumor subtypes, and individualized treatment. Recently, soft subspace clustering was introduced as an accurate alternative to conventional techniques. This practice has proven effective for high dimensional data, especially for microarray gene expressions. In this review, the basis of weighted dimensional space and different approaches to soft subspace clustering are described. Since most of the models are parameterized, the application of consensus clustering has been identified as a new research direction that is capable of turning the difficulty with parameter selection to an advantage of increasing diversity within an ensemble.

INTRODUCTION

Microarray technology has revolutionized biological and medical research. It becomes a central tool for examining gene expression profiles of a multitude of cells and tissues simultaneously. This innovation provides new opportunities to investigate and understand human disease. Gene expression data obtained from microarray experiments has inspired several novel applications, including the identification of differentially expressed genes for further molecular studies or

DOI: 10.4018/978-1-4666-4936-1.ch006

drug therapy response (Ramaswamy et al, 2003; Tusher et al, 2001; Wallqvist et al, 2002), and the creation of classification systems for improved cancer diagnosis (Cleator and Ashworth, 2004; Spang, 2003). Another typical analysis is to reveal natural structures and identify interesting patterns in the underlying microarray data (Jiang et al, 2004). The cluster analysis of biological samples using microarray data has been widely recognized as a standard practice in biological, clinical and toxicological studies (Carkacioglu et al, 2010; Chen et al, 2011; Golub et al, 1999). In particular to cancer research, it has become almost routine to create gene expression profiles, which can discriminate patients into good and poor prognosis groups, and identify possible tumor subtypes. This analysis offers a useful basis for individualized treatment of disease.

A variety of clustering algorithms and cluster ensemble methods are usually employed for the analysis of gene expression data (Iam-On et al, 2010). Initially, traditional algorithms such as *k*-means (McQueen, 1967) and agglomerative hierarchical clustering (Han and Kamber, 2000) have proven useful for identifying biologically relevant clusters of tissue samples and genes. In response to the challenges of high-dimensional data, especially in microarray gene expression data analysis, the practice of subspace clustering or biclustering (Cheng and Church, 2000; Gu and Liu, 2008; Prelic et al, 2006; Tanay et al, 2002; Tseng and Wong, 2005) has recently emerged as a new and effective alternative to any standard technique. Generally, cluster detection is based on a distance or proximity measure between objects of interest. However, with high-dimensional data, meaningful clusters cannot be easily identified as the distances between data objects are increasingly indifferent as dimensionality increases (Boongoen and Shen, 2010). In order to disclose patterns obscured by irrelevant dimensions, a global feature selection or reduction method, e.g., Principle Components Analysis (PCA; Joliffe, 1986), is effective only to a certain extent. Such a technique fails to detect in each dimension, locally varying relevance for distinct object groups. As a result, many different subspace clustering algorithms have been proposed with the common objective of discovering locally relevant dimensions per cluster (Boongoen et al, 2011; Kriegel et al, 2009). With the example in Figure 1 that represents different clusters of n objects $(x_1, x_2, ..., x_n)$ in d dimensions $(f_1, f_2, ...,$

Figure 1. Illustration of three different clusters: Cluster1 in a full dimensional space, Cluster2 and Cluster3 in distinct subspaces or subsets of the original dimensions