Chapter III

Kernel Clustering for Knowledge Discovery in Clinical Microarray Data Analysis

Nathalie L. M. M. Pochet, Katholieke Universiteit Leuven, Belgium

Fabian Ojeda, Katholieke Universiteit Leuven, Belgium

Frank De Smet, Katholieke Universiteit Leuven, Belgium & National Alliance of Christian Mutualities, Belgium

Tijl De Bie, Katholieke Universiteit Leuven, Belgium

Johan A. K. Suykens, Katholieke Universiteit Leuven, Belgium

Bart L. R. De Moor, Katholieke Universiteit Leuven, Belgium

Abstract

Clustering techniques like k-means and hierarchical clustering have shown to be useful when applied to microarray data for the identification of clinical classes, for example, in oncology. This chapter discusses the application of nonlinear techniques like kernel k-means and spectral clustering, which are based on kernel functions like linear and radial basis function (RBF) kernels. External validation techniques (e.g., the Rand index and the adjusted Rand index) can immediately be applied to these methods for the assessment of clustering results. Internal validation methods like the global silhouette index, the distortion score, and the Calinski-Harabasz index (F-statistic), which have been commonly used in the input space, are reformulated in this chapter for usage in a kernel-induced feature space.

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

Microarrays are a recent technology that allows for determining the expression levels of thousands of genes simultaneously. One important application area of this technology is clinical oncology. Parallel measurements of these expression levels result in data vectors that contain thousands of values, which are called expression patterns. A microarray consists of a reproducible pattern of several DNA probes attached to a small solid support. Labeled cDNA, prepared from extracted mRNA, is hybridized with the complementary DNA probes attached to the microarray. The hybridizations are measured by means of a laser scanner and transformed quantitatively. Two important types of microarrays are cDNA microarrays and oligonucleotide arrays. cDNA microarrays consist of about 10,000 known cDNA (obtained after PCR amplification) that are spotted in an ordered matrix on a glass slide. Oligonucleotide arrays (or DNA chips) are constructed by the synthesis of oligonucleotides on silicium chips. Figure 1 gives a schematic overview of an experiment with the cDNA technology. Both technologies have specific characteristics that will not be discussed here.

When studying, for example, tumor tissues with microarrays, the challenge mainly lies in the analysis of the experiments in order to obtain relevant clinical information. Most of the techniques that have been widely used for analyzing microarrays require some preprocessing stage such as gene selection, filtering, or dimensionality reduction, among others. These methods cannot directly deal with high-dimensional data vectors. Moreover, these are methods that are specifically designed to deal with the particular challenges posed by gene expression data and thus they do not provide a more general framework that can be easily extended to other kinds of data. For this purpose, methods and algorithms capable of handling high-dimensional data vectors and that are capable of working under a minimal set of assumptions are required. The chapter by Jean-Philippe Vert in this book focuses on the classification of high-dimensional data, while this chapter elaborates on the cluster analysis of these high-dimensional data.

Clustering techniques are generally applied to microarray data for the identification of clinical classes, which could allow for refining clinical management. Cluster analysis of entire microarray experiments (expression patterns from patients or tissues) allows for the discovery of possibly unknown diagnostic categories without knowing the properties of these classes in advance. These clusters could form the basis of new diagnostic schemes in which the different categories contain patients with less clinical variability.

Clustering microarray experiments have already shown to be useful in a large number of cancer studies. Alon et al. (1999), for example, separated cancerous colon tissues from noncancerous colon tissues by applying two-way clustering. The distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) has been rediscovered by using self-organizing maps (SOM) by Golub et al. (1999). By using hierarchical clustering, van ’t Veer et al. (2002) were able to distinguish between the presence (poor prognosis) and the absence (good prognosis) of distant subclinical metastases in breast cancer patients where the histopathological examination did not show tumor cells in local lymph nodes at diagnosis (lymph node negative).

For this purpose, methods such as the classical $k$-means clustering and hierarchical clustering are commonly used (Bolshakova, Azuaje, & Cunningham, 2005; Handl, Knowles, & Kell, 2005). These methods are based on simple distance or similarity measures (e.g., the
An Approach for the Development of Animal Tracking System


www.igi-global.com/article/an-approach-for-the-development-of-animal-tracking-system/201461?camid=4v1a