Parameterless Clustering Techniques for Gene Expression Analysis

Vincent S. Tseng, National Cheng Kung University, Taiwan
Ching-Pin Kao, National Cheng Kung University, Taiwan

Abstract

In recent years, clustering analysis has even become a valuable and useful tool for in-silico analysis of microarray or gene expression data. Although a number of clustering methods have been proposed, they are confronted with difficulties in meeting the requirements of automation, high quality, and high efficiency at the same time. In this chapter, we discuss the issue of parameterless clustering technique for gene expression analysis. We introduce two novel, parameterless and efficient clustering methods that fit for analysis of gene expression data. The unique feature of our methods is they incorporate the validation techniques into the clustering process so that high quality results can be obtained. Through experimental evaluation, these methods are shown to outperform other clustering methods greatly in terms of clustering quality, efficiency, and automation on both of synthetic and real data sets.

Copyright © 2006, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

Clustering analysis has been applied in a wide variety of fields such as biology, medicine, psychology, economics, sociology, and astrophysics. The main goal of clustering analysis is to partition a given set of objects into homogeneous groups based on their features such that objects within a group are more similar to each other and more different from those in other groups (Chen, Han, & Yu, 1996). Clustering groups the genes into biologically relevant clusters with similar expression patterns. The genes clustered together tend to be functionally related, hence clustering can reveal the co-expression of genes which were previously uncharacterized. In recent years, clustering analysis has even become a valuable and useful tool for in-silico analysis of microarray or gene expression data (Eisen et al., 1998; Alon et al., 1999; Ben-Dor & Yakhini, 1999; Tamayo et al., 1999). For example, Eisen et al. (1998) applied a variant of hierarchical clustering to identify groups of co-expressed yeast genes. Alon et al. (1999) used a two-way clustering technique to detect groups of correlated genes and tissues. Self-organizing maps were used by Tamayo et al. (1999) to identify clusters in the yeast cell cycle data set and human hematopoietic differentiation data set.

Thoroughly and extensive overviews of clustering algorithms are given by Aldenderfer and Blashfield (1984) and Jain and Dubes (1998). Although a number of clustering methods have been studied in the literature (Carpenter & Grossberg, 1987; Kohonen, 1990; Zhang, Ramakrishnan, & Livny, 1996; Chen, Han, & Yu, 1996; Guha, Rastogi, & Shim, 1998; Guha, Rastogi, & Shim, 1999), they are not satisfactory in terms of: (1) automation, (2) quality, and (3) efficiency. First, most clustering algorithms request users to specify some parameters. In real applications, however, it is hard for biologists to determine the suitable parameters manually. Thus an automated clustering method is required. Second, most clustering algorithms aim to produce the clustering results based on the input parameters and their own criterions. Hence, they are incapable of producing optimal clustering result. Third, the existing clustering algorithms may not perform well when the optimal or near-optimal clustering result is enforced from the universal criterions.

On the other hand, a variety of clustering validation measures are applied to evaluate the validity of the clustering results, the suitability of parameters, or the reliability of clustering algorithms. A good overview of clustering validation can be found in the book “Cluster Analysis” (Aldenderfer & Blashfield, 1984), in which numerous validation index are considered, like DB-index (Davies & Bouldin, 1979; Aldenderfer & Blashfield, 1984), Simple matching coefficient (Aldenderfer & Blashfield, 1984; Jain & Dubes, 1998), Jaccard coefficient (Aldenderfer & Blashfield, 1984; Jain & Dubes, 1998), Hubert’s Γ statistic (Aldenderfer & Blashfield, 1984; Jain & Dubes, 1998), etc. There also exist several other measures, like ANOVA (Kerr & Churchill, 2001), FOM (Yeung, Haynor, & Ruzzo, 2001), VCV (Hathaway & Bezdek, 2003). Nevertheless, the roles of them are placed only on the phase of “post-validation”. The study on how to integrate validation techniques with clustering methods tightly for improving the clustering quality has been absent.
Related Content

Mathematical Models of Dynamic Behavior of Individual Neural Networks of Central Nervous System
Dimitra-Despoina Pagania, Adam Adamopoulos and Spiridon Likothanassis (2012). *International Journal of Systems Biology and Biomedical Technologies* (pp. 48-65). [www.igi-global.com/article/mathematical-models-dynamic-behavior-individual/67106?camid=4v1a](www.igi-global.com/article/mathematical-models-dynamic-behavior-individual/67106?camid=4v1a)

Towards Optimal Microarray Universal Reference Sample Designs: An In-Silico Optimization Approach

A Modified High Speed Hopfield Neural Model for Perfect Calculation of Magnetic Resonance Spectroscopy
Hazem El-Bakry and Nikos Mastorakis (2010). *Biocomputation and Biomedical Informatics: Case Studies and Applications* (pp. 158-177). [www.igi-global.com/chapter/modified-high-speed-hopfield-neural/39610?camid=4v1a](www.igi-global.com/chapter/modified-high-speed-hopfield-neural/39610?camid=4v1a)