A Nearest Neighbor Clustering Algorithm for Gene Expression Data based on Iterative Sampling

Anca Ralescu and Waibhav Tembe
ECECS Department, University of Cincinnati, ML 0030, Cincinnati, Oh 45221
e-mail: Anca.Ralescu@uc.edu, wtembe@ececs.uc.edu

Introduction

Recently, the advent of the gene chip, also known as microarray analysis, has brought about the need to integrate new computational approaches in the processing and interpretation of the data provided by the gene chip. As a first step in processing such data clustering algorithms have been developed and/or adopted [1, ,]. High data dimensionality and volume call for efficient, yet biologically meaningful approaches so as to support and possibly guide further experimental work.

Algorithm Outline

We view the problem as a problem of vector clustering, where each vector represents the expression data for a gene across different experiments. The algorithm is based on calculating a matrix, D, where N is the number of genes to be clustered; D(i,j) is a measure of similarity between genes i and j. Clusters are formed recursively by adding genes according to the nearest neighbor criterion. Several issues arise in forming clusters this way, including: (1) detection of singletons and (2) possible merging of detected clusters or of singletons to existing clusters.
To cope with the large number of vectors (genes) the algorithm iteratively samples the genes and in each iteration detects new or updates, existing clusters. The algorithm ensures that the genes output as singletons (a singleton is a cluster of size 1, containing one gene only) at step i are considered in the sampling for subsequent steps. The iteration ends when all the genes in the data set have been considered at least once, or for a sufficiently large number of successive iterations clusters cease to be updated.

Another objective of the proposed approach is to enable incremental clustering in a distributed environment whereby clusters obtained from different data sets can be merged in a way compatible with clustering of the combined data sets.

The Data Set

The final goal is to use this algorithm for the yeast data set (Saccharomyces cervisiae), one of the publicly available data sets provided for this conference. At the current stage of development, and for the purpose of testing the algorithm, we have used a synthetic data set which closely resembles the real data set. We start with a set of distinct and independently generated vectors, V. For vector , a random number of vectors that should fall in the same cluster are generated by random perturbation of v. The collection of vectors generated this way makes up the synthetic data set.

Results

Extensive simulations show that results on the synthetic data set described above are very promising, with a near perfect clustering (clusters for the synthetic data are already known). The final results hinge on ongoing work on solving the cluster merging and singleton assignment (or non-assignment) to existing clusters.

Conclusion

A low complexity, of the order for cluster detection, not including the cluster merging step, transparent, flexible, clustering algorithm is proposed. Initial results on synthetic data are very promising. Future work include cluster analysis, summarization and prototyping.

References:

1. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein. Cluster Analysis and Display of Genome-wide Expression Patterns. PNAS Vol. 95, 14863-14868, December 1998.
2. E. Hartuv, A. O. Schmitt, J. Lange, S. Meyer-Ewert, H. Lehrach, R. Shamir. An Algorithm for Clustering cDNA Fingerprints. Genomics 66 249-256 (2000).
3. P. T. Spellman et al.. Comprehensive Identification of Cell Cylce-regulated genes of the yeast Saccharomyces cervisiae by Microarray Hybridization. Molecular Biology of the Cell, Vol. 9, 3273-3297, December 1998.