Mining Gene Expression Datasets using Density-based Clustering [*]

Seokkyung Chung, Jongeun Jun, Dennis McLeod
Department of Computer Science
and Integrated Media System Center
University of Southern California
Los Angeles, California 90089–0781, USA
[seokkyuc, jongeunj, mcleod]@usc.edu

ABSTRACT

Given the recent advancement of microarray technologies, we present a density-based clustering approach for the purpose of co-expressed gene cluster identification. The underlying hypothesis is that a set of co-expressed gene clusters can be used to reveal a common biological function. By addressing the strengths and limitations of previous density-based clustering approaches, we present a novel clustering algorithm that utilizes a neighborhood defined by k-nearest neighbors. Experimental results indicate that the proposed method identifies biologically meaningful and co-expressed gene clusters.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering

General Terms

Algorithms

Keywords

Density-based Clustering, Gene Expression Analysis, Microarray Analysis

1. INTRODUCTION

With the recent advancement of DNA microarray technologies, the expression levels of thousands of genes can be measured simultaneously. The obtained data are usually organized as a matrix where the columns represent genes (usually genes of the whole genome), and the rows correspond to the samples (e.g. various tissues, experimental conditions, or time points). Given this rich amount of gene expression data, it is essential to extract hidden knowledge from this matrix.

One of the key steps in gene expression analysis is to perform the clustering of genes that show similar patterns. By identifying a set of gene clusters, we can hypothesize that the genes clustered together tend to be functionally related. Thus, gene expression clustering may be useful in identifying mechanisms of gene regulation and interaction, which can be used to understand the function of a cell.

Since gene expression data consist of measurements across various conditions (or time points), they are characterized by multi-dimensional, huge size in terms of volume, and noisy data. Thus, clustering algorithms must be able to address and exploit such features of the datasets. Recent database mining research has proposed density-based clustering algorithms, which are relevant for multi-dimensional noisy datasets. By addressing the limitations of previous density-based clustering methods, we present a novel KNN (k-nearest neighbor) density estimation clustering algorithm that is relevant for producing co-expressed gene clusters.

In this paper, we mainly focus on time-course gene expression data (i.e., expression levels of genes that are monitored during some time interval). In particular, we utilize the yeast cell cycle dataset introduced in Spellman *et al.* [3].

2. PROPOSED ALGORITHM

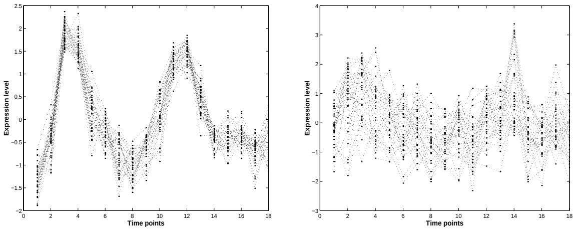
Conventional density-based clustering starts by estimating the density for each point in order to identify core, border and noise points. A core point is referred to as a point whose density is greater than a user-defined threshold. Similarly, a noise point is referred to as a point whose density is less than a user-defined threshold. Noise points are usually discarded in the clustering process. A non-core, non-noise point is considered as a border point. Hence, clusters can be defined as dense regions (i.e., a set of core points), and each dense region is separated from one another by low density regions (i.e., a set of border points).

By incorporating the ideas of density-based clustering, our clustering algorithm proceeds in three phases: (1) density estimation for each gene; (2) rough cluster identification using core genes (i.e., core points); (3) cluster refinement using border genes (i.e., border points). In this short paper, we only sketch the main idea of our algorithm. For details, please refer to Chung *et al.* [1].

For density estimation, the proposed algorithm is mainly

[*] This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

CIKM'04, November 8–13, 2004, Washington, DC, USA.
ACM 1-58113-874-1/04/0011.



(a) High density gene (b) Low density gene

Figure 1: Plot of top k-nearest neighbors for a high density gene and a low density gene (when k=30)

focused on KNN (k-nearest neighbor) density estimation. That is, density of a gene, x, is defined by the sum of similarities between k-nearest neighbors to x. Figure 1 illustrates the intuition behind this approach. As shown, with a high density gene (Figure 1(a)), the sum of similarity between nearest neighbors to itself is relatively higher than a low density gene (Figure 1(b)). In addition, since the overall shape of gene expression patterns is more important than magnititude in gene expression datasets, we use the Pearson correlation coefficient as the similarity metric among two genes.

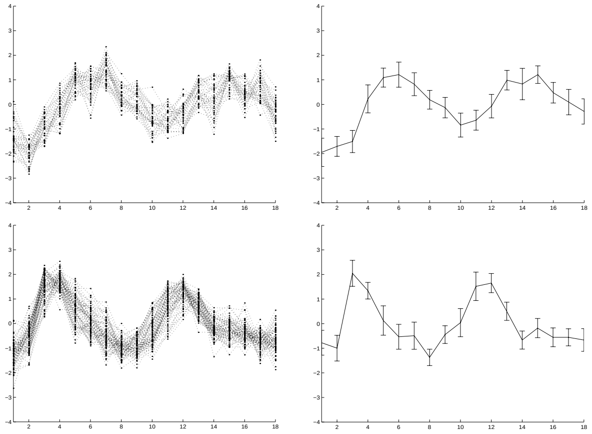
Since a core gene has high density, it is expected to locate well inside the cluster. Thus, instead of performing clustering on entire datasets, conducting clustering on core genes can produce a rough cluster structure. Since border and noise genes are excluded in the rough cluster identification step, each cluster is expected to be well separated from other clusters. Once the skeleton of a cluster structure is identified, border genes are used to refine the cluster structure by assigning those border genes to the most relevant cluster.

For some biological processes (e.g., the cell cycle), expression relationships may be revealed at different time points. In the presence of such a time-shift, Pearson's correlation is limited in capturing the relationship between two expression profiles. Moreover, some mechanisms (e.g., negative feedback loops) can limit the number of expressed genes based on the principle of efficiency. Thus, the expression relationships among genes along the same biological pathway may be partially revealed in a single microarray experiment. Therefore, in order to address the two problems, the proposed clustering algorithm exploits neighborhood-based clustering [2].

To utilize spatial index structures for efficient density estimation, we conduct dimensionality reduction based on Singular Value Decomposition (SVD) [1]. Although SVD has been utilized in gene expression clustering research, our main purpose is different in that the main goal of the previous approaches was to preprocess the data before clustering while our main aim is to efficiently support similarity search in the truncated SVD space.

3. EXPERIMENTAL RESULTS

We applied our algorithm to Spellman's dataset. Figure 2 plots sample clusters where the x-axis and y-axis represents time points and expression values, respectively. Figure 2(a) shows sample core clusters, and Figure 2(b) illustrates the corresponding coherent patterns that characterize a trend of expression levels of genes within a cluster. A coherent pattern of a cluster is defined by a medoid of the cluster.



(a) Core clusters (b) Coherent patterns

Figure 2: Sample examples of core clusters and the corresponding coherent expression patterns

The first cluster is mainly composed of genes that are involved in the assembly and arrangement of cell structures. For example, YBR009C, YNL030W, and YDR224C (whose biological function corresponds to chromatin assembly or disassembly) are classified into the first cluster. YMR307W (whose function is cell wall organization and biogenesis) and YCL063W (whose function is vacuole inheritance) are also classified into the first cluster. YNL339C, YLR467W, YAL007C, and YAL014C (whose biological process is telomerase independent telomere maintenance) are detected as the second cluster. In addition, YOR033C and YDR097C (whose biological function corresponds to mismatch repair) are also classified together.

4. CONCLUSION AND FUTURE WORK

We presented a mining framework that is useful in microarray data analysis. An experimental prototype system has been developed, implemented, and tested to demonstrate the effectiveness of the proposed model. In order to identify co-expressed genes in a yeast cell cycle dataset, we developed the clustering algorithm based on KNN density estimation. The proposed clustering algorithm successfully identified co-expressed clusters. In the future, we plan to incorporate biological annotation into the clustering process.

5. REFERENCES

[1] S. Chung, J. Jun, and D. McLeod. Mining gene expression datasets using density-based clustering. In *USC/IMSC Technical Report, IMSC-04-002*, 2004.

[2] S. Chung, and D. McLeod. Dynamic topic mining from news stream data. In *Proceedings of ODBASE*, 2003.

[3] P. T. Spellman *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces Cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273-3297, 1998.