

Integration of Cluster Ensemble and EM based Text Mining for Microarray Gene Cluster Identification and Annotation

Xiaohua Hu
College of IST, Drexel Univ.
3141 Chestnut Street
Philadelphia, PA 19104, USA
thu@ischool.drexel.edu

Xiaodan Zhang
College of IST, Drexel Univ.
3141 Chestnut Street
Philadelphia, PA 19104, USA
xzhang@ischool.drexel.edu

Xiaohua Zhou
College of IST, Drexel Univ.
3141 Chestnut Street
Philadelphia, PA 19104, USA
Xiaohua.zhou@drexel.edu

ABSTRACT

In this paper, we design and develop a unified system GE-Miner (Gene Expression Miner) to integrate cluster ensemble, text clustering and multi document summarization and provide an environment for comprehensive gene expression data analysis. We present a novel cluster ensemble approach to generate high quality gene cluster. In our text summarization module, given a gene cluster, our Expectation Maximization (EM) based algorithm can automatically identify subtopics and extract most probable terms for each topic. Then, the extracted top k topical terms from each subtopic are combined to form the biological explanation of each gene cluster. Experimental results demonstrate that our system can obtain high quality clusters and provide informative key terms for the gene clusters.

Categories and Subject Descriptors

D.3.3 [Computer Applications]: LIFE AND MEDICAL SCIENCES – *Biology and genetics.*

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Experimentation

Keywords

Gene Cluster Identification, Gene Cluster Annotation, Cluster Ensemble, Text Mining, EM, Summarization

1. INTRODUCTION

Huge amounts of gene expression data have been generated as a result of the Human Genomic project, which creates a need and challenge for data mining. Clustering algorithms are used as essential tools to analyze gene expression data sets and provide valuable insight on various aspects of the genetic machinery such as identifying the functionality of genes, finding out what genes are co-regulated, distinguishing the important genes between abnormal tissue and normal tissues, etc. Generating high quality gene clusters and identifying underlying biological mechanism of the gene cluster are the ultimate goal of clustering gene expression analysis. But there are some drawbacks of these

approaches. To get high quality cluster results, these approaches rely on choosing the best cluster algorithm whose design biases and assumptions meet the underlying distribution of the data set. Otherwise, the results will be poor if the assumptions are violated in a data set. On the other hand, clustering indeed reveals potential meaningful relationships among genes, but cannot explain the underlying biological mechanisms. Another drawback is that the clustering quality and cluster interpretation are treated as two isolated research problems and are studied separately. But cluster quality and cluster interpretation are closely related and must be addressed in a coherent and unified way. Based on this consideration, this paper explores the first step toward dealing with these issues. We design and develop a unified system GE-Miner (Gene Expression Miner) to address these challenging issues in a principled and general manner by integrating cluster ensemble and text summarization and provide an environment for comprehensive gene expression data analysis. The task of establishing a unifying framework for comprehensive gene expression analysis is accomplished in three steps: 1) Cluster Ensemble: building a cluster ensemble method to combine the clustering results from various clustering algorithms in order to obtain high quality and robust results; 2) Data Integration Server: developing an extendable data integration server to gather related textual resource from various databases of the genes; 3) Textual Summarization: integrating biomedical literature mining in gene expression analysis to provide informative biological explanation of the gene clusters.

2. GE-MINER ARCHITECTURE

We develop a comprehensive gene expression mining system **GE-Miner** (Gene Expression Miner) as shown in Figure 1 geared precisely for this task, helping a biologist in cross-referencing experimental and analytical results obtained from microarray experiments and provide concise and meaningful biological explanation of the gene clusters. **GE-Miner** aims to summarize the biomedical knowledge for genes on a genome-wide scale, generates relevant summaries from relevant biological literatures and summarizes biological information about a group of genes in a concise and coherent manner. It has an open-architecture and can easily add a wrapper if new data sources become available, without affecting the rest of the system. The components of GE-Miner are described in details in the following sections. For details, please refer to our former work [1].

3. CLUSTER ENSEMBLE

At the first step, various clustering algorithms are run against the same data sets to generate clustering results. Then, these

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'06, November 5-11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011...\$5.00.

clustering results are combined by an auto-associative additive system based on the distance matrix of graph clustering. In our approach, a distance matrix is first constructed based on the cluster results from each individual clustering algorithm; these similarity matrices are combined to form a master distance matrix. Then a similarity graph is constructed from the master distance matrix and a graph-based partitioning algorithm is applied to the graph for the final clustering results. Graph-based clustering uses various kinds of geometric structure or graphs for analyzing data. Different graphs reflect various local structure or inherent visual characteristic in the data set. Clustering divides the graph into connected components by identifying and deleting inconsistent edges, and each subgraph consisting of connected components refers to a cluster.

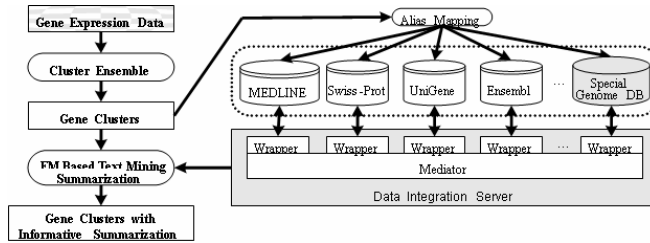


Fig. 1. The Architecture of Gene Expression Miner (GE-Miner)

4. EM-BASED INFORMATIVE TEXTUAL SUMMARIZATION OF GENE CLUSTERS

Our approach takes each gene cluster as a major topic. Each major topic is summarized as several subtopics. Each subtopic contains certain number of documents and terms. Therefore, the related document set of a gene cluster is clustered into a few document term clusters; each document term cluster corresponds to a subtopic of the gene cluster. Our rationale of summarization is that the document set of a gene cluster contains various biological subtopics. An Expectation Maximization (EM) based multi document summarization algorithm is applied to summarize each gene cluster separately. The EM based algorithm automatically extracts most probable terms and documents for each sub topics. Finally the extracted top k terms from each subtopic are combined to form the biological explanation of each gene cluster. Compared to other methods [2], we have several advantages. First, we assume different text section of a document plays different role for estimating the probability of a term belonging to a subtopic when it's known that some documents belong to this subtopic. A two component mixture model is accordingly developed to estimate term probability belonging to a subtopic. Thus, it improves the classification and summarization results of EM when treating terms from different section of text differently. Second, for term probability smoothing, we import a mixture weights parameter α to control the portion of how much a term probability is distributed according to subtopic and collection frequency, which serves as lower ranking terms appear in most of documents. Third, the presentation of a document is also different. While they use unigram to represent a document, we use medical concepts to represent a document, which help catch more meaningful phrases. The medical concepts are extracted by part of speech tagging using UMLS ontology. As for space, we only list important formulas (for others, please refer to [2]). The complete log likelihood is expressed as follows:

$$l_c(\theta | D, z) = \log P(\theta) + \sum_{d_i \in D} \sum_{j=1}^{|C|} z_{ij} \log(P(c_j | \theta) P(d_i | c_j)) \quad (1)$$

where D represents the document collection, d represents document, C is class label, $Z\{1,0\}$ indicates whether the document belong to the given class C .

$$\hat{\theta}_{t_m | c_j} = P(t_m | c_j; \hat{\theta}) = \lambda P(t_m | Title, C_j; \hat{\theta}) + (1 - \lambda) P(t_m | Abstract, C_j; \hat{\theta}) \quad (2)$$

$$\hat{\theta}_{t | c_j} = \alpha P(t | C_j; \hat{\theta}) + (1 - \alpha) P(t | collection; \hat{\theta}) \quad (3)$$

$$\lambda^{(n+1)} = \frac{1}{|T|} \frac{1}{K} * \quad (4)$$

$$\sum_{t_m \in T} \frac{\lambda^{(n)} P(t_m | Title, C_j; \hat{\theta})}{\lambda^{(n)} P(t_m | Title, C_j; \hat{\theta}) + (1 - \lambda^{(n)}) P(t_m | Abstract, C_j; \hat{\theta})}$$

First, we assume a term is either generated from a class model or from collection model which is interpolated by α (see equation (3)); second, we assume a term from different section of the text plays different role on the clustering process which is interpolated by λ (see (2)). Equation (4) is the updating formula for parameter λ , where K indicates the number of subtopics.

5. CONCLUSION

In this paper we present a novel system GE-Miner for comprehensive gene expression analysis. Moreover, we provide a novel EM based multi document summarization method. Especially, we utilize a mixture model of terms from different text sections of documents to estimate term subtopic probability and applies collection frequency smoothing method to remove background noise. Experiment results show that the trained coefficient λ of the mixture model constantly get a stable value about 0.1 which indicates that title terms contribute more than abstract terms to subtopics. In practice, the algorithm can automatically identify sub topics and assign most probable terms to each subtopic, which are combined to form the biological explanation of each gene cluster.

6. ACKNOWLEDGMENTS

Hu's work is supported partially by the NSF Career grant IIS 0448023 and NSF CCF 0514679 and PA Dept of Health Tobacco Formula Grants.

7. REFERENCE

- [1] Hu X., *Integration of Cluster Ensemble and Text Summarization for Gene Expression Analysis*, in Proceedings of IEEE 2004 Symposium on Bioinformatics and Bioengineering, 251-259, May 19-21, 2004, Taiwan (IEEE BIBE 2004)
- [2] Kamal Nigam, Andrew McCallum, Sebastian Thrun and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3). pp. 103-134. 2000.