Chapter 9
An Integrated Framework for Fuzzy Classification and Analysis of Gene Expression Data

Mohammad Khabbaz
University of Calgary, Canada

Keivan Kianmehr
University of Calgary, Canada

Mohammad Alshalalfa
University of Calgary, Canada

Reda Alhajj
University of Calgary, Canada

ABSTRACT

This chapter takes advantage of using fuzzy classifier rules to capture the correlations between genes. The main motivation to conduct this study is that a fuzzy classifier rule is essentially an “if-then” rule that contains linguistic terms to represent the feature values. This representation of a rule that demonstrates the correlations among the genes is very simple to understand and interpret for domain experts. In this proposed gene selection procedure, instead of measuring the effectiveness of every single gene for building the classifier model, the authors incorporate the impotence of a gene correlation with other existing genes in the process of gene selection. That is, a gene is rejected if it is not in a significant correlation with other genes in the dataset. Furthermore, in order to improve the reliability of this approach, the process is repeated several times in these experiments, and the genes reported as the result are the genes selected in most experiments. This chapter reports test results on five datasets and analyzes the achieved results from biological perspective.

DOI: 10.4018/978-1-60566-717-1.ch009

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION

Gene expression analysis is the use of quantitative RNA measurements of gene expression in order to characterize biological processes and clarify the mechanisms of gene transcription. The microarray technology makes it possible to monitor the expression levels of tens of thousands of genes in parallel. It measures the relative RNA levels of genes and produces high dimensional dataset, which is in turn a highly informative source for gene expression analysis. The microarray data from a series of N experiments may be represented as N×M gene expression matrix in which each of the N rows consists of an M-element expression vector. The latter vector represents the gene expression level of the genes for a single experiment.

Table 1 displays an example matrix for gene expression data, where the rows denote different samples or conditions (such as the same cell type among different samples), while the columns denote genes; F [Sample 3, G₄] denotes the quantitative expression of gene G₄ in Sample 3.

These samples may correlate with different time points taken during a biological process or with different tissue types, such as normal cells and cancer cells Aas (2001). To this end, data mining techniques have been widely applied to gene expression data for conducting analysis in gaining insight into the functional behavior of genes as well as to correlate structural information with functional information. The main methods of microarray data analysis include Piatetsky-Shapiro & Tamayo (2003):

- **Cluster Analysis**: This can be performed to identify genes that reveal similar functionalities under a number of experimental conditions.
- **Discriminant Analysis (Classification and prediction)**: This can potentially yield useful diagnostic tools for classifying samples on the basis of their gene expression patterns.
- **Gene Selection**: This can yield a small fraction of genes that are informative in discriminant analysis problems.

Although microarray provides considerable flexibility for gene expression analysis, there are several general issues in using microarray data. First, the datasets are usually complex and noisy. The noise and complexity in experimental protocols strongly limit data integration. Another important issue is the fact that currently available datasets typically contain fewer than one hundred instances, though each instance quantifies the expression levels of several thousands of genes (i.e., high dimensionality). For instance in gene expression classification, due to the high dimensionality and the small sample size of the experimental data, it is often possible to find a large number of classifiers that can separate the training data perfectly, but their diagnostic ac-

<table>
<thead>
<tr>
<th>Sample</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>52</td>
<td>39</td>
<td>59</td>
<td>16</td>
</tr>
<tr>
<td>Sample2</td>
<td>57</td>
<td>32</td>
<td>64</td>
<td>21</td>
</tr>
<tr>
<td>Sample3</td>
<td>30</td>
<td>76</td>
<td>61</td>
<td>65</td>
</tr>
<tr>
<td>Sample4</td>
<td>52</td>
<td>39</td>
<td>59</td>
<td>16</td>
</tr>
<tr>
<td>Sample5</td>
<td>18</td>
<td>79</td>
<td>56</td>
<td>64</td>
</tr>
<tr>
<td>Sample6</td>
<td>65</td>
<td>81</td>
<td>59</td>
<td>70</td>
</tr>
<tr>
<td>Sample7</td>
<td>41</td>
<td>83</td>
<td>21</td>
<td>23</td>
</tr>
</tbody>
</table>
Related Content

Current Issues and Future Analysis in Text Mining for Information Security Applications
www.igi-global.com/chapter/current-issues-future-analysis-text/21752?camid=4v1a

Data Mining Methods for Crude Oil Market Analysis and Forecast
Jue Wang, Wei Xu, Xun Zhang, Yejing Bao, Ye Pang and Shouyang Wang (2010). Data Mining in Public and Private Sectors: Organizational and Government Applications (pp. 184-203).
www.igi-global.com/chapter/data-mining-methods-crude-oil/44289?camid=4v1a

Feature Selection for the Promoter Recognition and Prediction Problem
George Potamias and Alexandros Kanterakis (2007). International Journal of Data Warehousing and Mining (pp. 60-78).
www.igi-global.com/article/feature-selection-promoter-recognition-prediction/1790?camid=4v1a

Algebraic and Graphic Languages for OLAP Manipulations
www.igi-global.com/article/algebraic-graphic-languages-olap-manipulations/1798?camid=4v1a