Probability Based Most Informative Gene Selection From Microarray Data

Sunanda Das, Neotia Institute of Technology, Management and Science, Department of Computer Science and Engineering, Diamond Harbour, W. B., India

Asit Kumar Das, Indian Institute of Engineering Science and Technology, Shibpur, Department of Computer Science and Technology, Howrah, W. B., India

ABSTRACT

Microarray datasets have a wide application in bioinformatics research. Analysis to measure the expression level of thousands of genes of this kind of high-throughput data can help for finding the cause and subsequent treatment of any disease. There are many techniques in gene analysis to extract biologically relevant information from inconsistent and ambiguous data. In this paper, the concepts of functional dependency and closure of an attribute of database technology are used for finding the most important set of genes for cancer detection. Firstly, the method computes similarity factor between each pair of genes. Based on the similarity factors a set of gene dependency is formed from which closure set is obtained. Subsequently, conditional probability based interestingness measurements are used to determine the most informative gene for disease classification. The proposed method is applied on some publicly available cancerous gene expression dataset. The result shows the effectiveness and robustness of the algorithm.

KEYWORDS

Important Gene Set, Most Informative Gene Selection, Probability Factor, Similarity Based Gene Dependency

INTRODUCTION

Now, the rapid growth of databases is undergoing beyond the human’s capability for analysis. Microarray technology helps to measure the expression levels of thousands of genes on a genome-wide scale (Krane & Raymer, 2003; Verdik, 2004; Golub et al., 1999). Many investigations and research on microarray technology reveal that cancer is a disease which causes dynamic changes in the genome. This microarray technology helps to build cancer diagnosis, prognosis or prediction of classifiers.

Generally, various models for cancer prediction and rule generation have been proposed such as k-NNs (k-Nearest Neighbours), Naive Bayes (NB), ANNs (Artificial Neural Networks), SVMs (Support Vector Machines), GAs (Genetic Algorithms), DTs (Decision Trees), RSs (Rough Sets) et al (Rana & Lal, 2016). Many researchers show that Bayesian decision theoretic rough set model (Bhattacharya & Debnath, 2014; Setiawan, 2014) gives better result than other method. In this article, the theory of Jaccard Similarity is used to find dependence relationship among data and evaluate the importance of attributes. In databases, the knowledge discovery (Fayy et al., 1996) is the process of finding useful patterns or meaning in raw data. Today in the field of the medical domain, medical data mining creates great potential for exploring the hidden patterns. But still it cannot manage inconsistent information. So, it’s a crucial task in this field to find gene selection (Li & Zhang, 2006) and classifier construction. Identifying a few informative genes from thousands or tens of thousands...
of genes to construct accurate prediction models is not an easy task. In this paper, a method for the classification of cancer is presented based on gene expression profiles using single gene. Already few researchers show the accurate classification results obtained based on using gene pairs (Gordon, 2002; Geman et al., 2004). Using single gene simple prediction models can easily be obtained for accurate cancerous prediction. Recently, few researchers proposed a rough set based soft computing method to conduct cancer classification using single or double genes (Wang, X. & Gotoh, 2009; Wang & Tetko, 2005; Li & Wong, 2002). Here, a simple rule-based cancer prediction model is constructed for easy understanding of biologists and clinicians.

So, selection of feature subset properly is a vital task for classification problems. Consequently, many learning algorithms can be used to obtain a set of rules in IF-THEN form, from a decision table. Four microarray datasets is used for study, which has been widely used in cancer classification experiments.

THE PROPOSED ALGORITHM

The data studied here for the proposed method is represented in the form of decision table. One decision table can be represented as $S = (U, C, D)$, where $U$ is the set of samples, $C= \{g_1, g_2, \ldots, g_n\}$ is the condition attribute set and $D = \{d_1, d_2\}$ is the decision attribute set. Every gene expression data can be represented with the decision table which is shown in Table 1.

For any particular cancer disease, as all genes are not important, a relevance analysis of genes to select only the important genes is necessary. The overall proposed method is described in Figure 1.

Data Preprocessing

Now-a-days discrete values play vital role in knowledge representation than continuous attributes. As all learning algorithms are incapable to handle continuous attributes, different discretization techniques (Kurgan & Cios, 2004; Tsai et al., 2008; Butterworth & Simovici, 2004; Fayyad & Irani, 1993) are proposed by many authors. The chimerge discretization (Kerber, 1992) is performed here for the proposed method on some microarray data which are available at datam.i2r.a-star.edu.sg/datasets/krbd/.

Similarity Based Gene Dependency

Let $g_i$ and $g_j$ belongs to the sample set $\{s_{i_1}, s_{i_2}, \ldots s_{i_h}\}$ and $\{s_{j_1}, s_{j_2}, \ldots s_{j_k}\}$, respectively. Similarity factor $S_{g_i g_j}$ between two genes $g_i$ and $g_j$ are computed using equation (1).

$$S_{g_i g_j} = \frac{\big(\big(s_{i_1}, s_{i_2}, \ldots s_{i_h}\big) \cap \big(s_{j_1}, s_{j_2}, \ldots s_{j_k}\big)\big)}{\big(\big(s_{i_1}, s_{i_2}, \ldots s_{i_h}\big) \cup \big(s_{j_1}, s_{j_2}, \ldots s_{j_k}\big)\big)}$$

(1)

Table 1. Microarray dataset decision table

<table>
<thead>
<tr>
<th>Samples</th>
<th>Condition attributes (genes)</th>
<th>Decision attributes (classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gene 1</td>
<td>Gene 2</td>
</tr>
<tr>
<td>1</td>
<td>g(1, 1)</td>
<td>g(1, 2)</td>
</tr>
<tr>
<td>2</td>
<td>g(2, 1)</td>
<td>g(2, 2)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>m</td>
<td>g(m, 1)</td>
<td>g(m, 2)</td>
</tr>
</tbody>
</table>
Trusted and Trustworthy Information Technology
www.igi-global.com/chapter/trusted-and-trustworthy-information-technology/112885?camid=4v1a

Use of Online Social Networking and Academic Performance of Students
www.igi-global.com/chapter/use-of-online-social-networking-and-academic-performance-of-students/112682?camid=4v1a