Egyptian Informatics Journal (2011) **12**, 73–82

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Suite of decision tree-based classification algorithms on cancer gene expression data

Mohmad Badr Al Snousy [a], Hesham Mohamed El-Deeb [b,*], Khaled Badran [c], Ibrahim Ali Al Khlil [c]

[a] *Department of Computer Science, Sadat Academy for Management Science (SAMS), Egypt*
[b] *Department of Computer Science, Modern University for Technology and Information (M.T.I.), Egypt*
[c] *Department of Computer Science, Military Technical College (M.T.C.), Egypt*

Received 28 December 2010; accepted 5 April 2011
Available online 23 July 2011

KEYWORDS

DNA microarray;
Cancer;
Classification;
Decision trees;
Ensample decision tree;
Attribute selection

Abstract One of the major challenges in microarray analysis, especially in cancer gene expression profiles, is to determine genes or groups of genes that are highly expressed in cancer cells but not in normal cells. Supervised machine learning techniques are used with microarray datasets to build classification models that improve the diagnostic of different diseases. In this study, we compare the classification accuracy among nine decision tree methods; which are divided into two main categories; the first is single decision tree C4.5, CART, Decision Stump, Random Tree and REPTree. The second category is ensample decision tree such Bagging (C4.5 and REPTree), AdaBoost (C4.5 and REPTree), ADTree, and Random Forests. In addition to the previous comparative analyses, we evaluate the behaviors of these methods with/without applying attribute selection (A.S.) techniques such as Chi-square attribute selection and Gain Ratio attribute selection. Usually, the ensembles learning methods: bagging, boosting, and Random Forest; enhanced classification accuracy of single decision tree due to the natures of its mechanism which generate several classifiers from one dataset and vote for their classification decision. The values of enhancement fluctuate

* Corresponding author.
E-mail addresses: badr_senousy_arcoit@yahoo.com (M.B.A. Snousy), hmeldeeb14@yahoo.com (H.M. El-Deeb), khaledBadran@hotmail.com (K. Badran), ibrahim.alkhlil@gmail.com (I.A.A. Khlil).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.
doi:10.1016/j.eij.2011.04.003

Production and hosting by Elsevier

between (4.99–6.19%). In majority of datasets and classification methods, Gain ratio attribute selection slightly enhanced the classification accuracy (~1.05%) due to the concentration on the most promising genes having the effective information gain that discriminate the dataset. Also, Chi-square attributes evaluation for ensemble classifiers slightly decreased the classification accuracy due to the elimination of some informative genes.

1. Introduction

The genome ribonucleic acid (RNA) expression studies allow systematic approaches to understand the relationship between gene expression profiles and disease states or different developmental stages of a cell. Microarray analysis provides quantitative information about the whole transcription profile of cells that make possible drug and therapeutics improvement, disease diagnosis, and comprehensible basic cell biology.

A DNA microarray technique allows to simultaneously observing the expression levels of thousands of genes during significant biological processes and across collections of related samples [1].

The datasets from microarray analysis, that enables the measurement of molecular signatures of diverse cells, becomes an important application of data mining, artificial intelligence and machine learning techniques to provide bioinformatics knowledge. In practical, supervised machine learning techniques used with microarray datasets to build classification models that improve the diagnostic of different diseases which easy to interpreted [2,3].

1.1. Biological background

Cells are the fundamental working units of every living system. All the instructions needed to direct their actions are contained within the chemical deoxyribonucleic acid or shortly DNA. A DNA molecule is a double-stranded polymer composed of four basic molecular units namely nucleotides. The nitrogen bases include adenine (A), guanine (G), cytosine (C) and thymine (T). The genome provides a template for the synthesis of a variety of RNA molecules. The process of transcribing a gene's DNA sequence into RNA is called gene expression. A gene's expression level indicates the approximate number of copies that gene's RNA produced in a cell and it is correlated with the amount of the corresponding proteins made. This mechanism controls which genes are expressed in a cell and acts as a "volume control" that increases or decreases the level of expression of particular genes as necessary [4].

1.2. Microarray data format

A gene expression data set from a microarray experiment can be represented by a real-valued.

$$\text{Expression matrix} = \{G(i,j) | 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$$

where the columns $G = \{\vec{g_1}, \vec{g_2}, \ldots, \vec{g_m}\}$ form the expression patterns of genes, the rows $S = \{\vec{S_1}, \vec{S_2}, \ldots \vec{S_n}\}$.

An example of a gene expression microarray dataset for Leukemia is shown in Table 1. the table organizes data into m columns (genes) and n rows (samples) where m mostly varies from thousand to hundred thousand according to the accuracy of microarray image processing technique, while n is always less than 200 samples according to the previously collected datasets [5]. Category column presents the actual class of the sample. For the shown example AML stands for acute myeloid leukemia disease and ALL represents acute lymphoblastic.

Table 1 Microarray data decision table.

Samples	Attributes (genes)				Category
	Gene 1	Gene 2	...	Gene m	
1	G(1,1)	G(1,2)	...	G(1,m)	ALL
2	G(2,1)	G(2,2)	...	G(2,m)	ALL
...	...	...	...	...	ALL
...	...	...	...	...	AML
n	G(n,1)	G(n,2)	...	G(n,m)	AML

Our study provides a performance comparison of nine decision tree methods. The rest of this paper is organized as the follows. In Section 2, we present brief challenges that faced in cancer classification area. In Section 3, we provide problem definitions. In Section 4, we exploit decision tree and microarray classification. In Section 5, we discuss related works in this domain. In Section 6, we explore the methodologies used in this work. In Section 7, we describe experimental setup. In Section 8, we present results and analysis. In Section 9, we conclude the paper.

2. Cancer classification challenges

Gene classification as domain of research poses a new challenges due to its unique problem nature. First, challenge comes from the unique nature of the available gene expression dataset; where most of these datasets has sample size below 200, vs. thousands to hundred thousands of genes presented in each tuples. Second, only a few numbers of these (genes) presents relevant attributes to the investigated disease. Third, comes from the presence of noise (biological and technical) inherent in the dataset. Fourth challenge arises from the application area, for instance accuracy is an important criterion in cancer classification task, but it is not the only goal, in cancer domain we want to achieve, biological relevancy as well as classification accuracy.

3. Problem definition

There is no single classifier superior over the rest, for instance the classification accuracy is depend on the classification method, gene selection method, and dataset [7,8].

In this study we will use the notation provided by Ying Lu et al. [9].

Let $X_1, X_2, \ldots, X_m$ be random variables for genes G_1, $G_2, \ldots, G_m$ respectively, where X_i has domain $dom\ (X_i)$ which is the range of expression values for *gene* G_i.