Filter/Wrapper Methods for Gene Selection and Classification of Microarray Dataset

Norreddine Mekour, Norreddine Mekour, GeCoDe Laboratory, Department of Computer Science, Tahar Moulay University of Saida, Saida, Algeria
Reda Mohamed Hamou, GeCoDe Laboratory, Department of Computer Science, Tahar Moulay University of Saida, Saida, Algeria
Abdelmalek Amine, GeCoDe Laboratory, Department of Computer Science, Tahar Moulay University of Saida, Saida, Algeria

https://orcid.org/0000-0001-9327-7903

ABSTRACT

A wide variety of large-scale information has been made within the extraction of genomic information and the extraction of data. Problems addressed embody ordination sequencing, supermolecule structure modeling, or the reconstruction of biological process trees (phylogeny). These issues need collaboration between biologists and computer scientists as a result of the issues to be of nice recursive difficulties. One of the most modern problems that gene expression data is resolved is with feature selection. There are two general approaches for feature selection: filter approach and wrapper approach. In this article, the authors propose a new approach when combining the filter approach with method ranked information gain and a wrapper approach with the searching method of the genetic algorithm. In order to test their overall performance, an experimental study is presented based on two gene microarray datasets found in bioinformatics and biomedical domains leukemia, and the central nervous system (CNS). The classifier Decision tree (C4.5) is used for improving the classification performance. The results show that their approach selects genes for additional correct classification emphasizes the effectiveness of the chosen genes and its ability to filter the information from unsuitable genes.

KEYWORDS

Classification, Data Mining, Decision Tree, Feature Selection, Gene Expression, Genetic Algorithms, Knowledge Extraction

1. INTRODUCTION

The field of bioinformatics has for many years already attracted an excellent interest within the scientific community as a result of it disclose terribly made views for the understanding of biological phenomena. Problems addressed include genome sequencing, protein structure modeling, or the reconstruction of phylogenetic trees (phylogeny), attribute selection. These problems require collaboration between biologists and computer scientists because the problems to be dealt with often pose great algorithmic difficulties.

Nowadays, folks will get the expression datasets of thousands of genes at the same time victimization microarray technology. One among the vital fields in victimization these organic phenomenon datasets is to classify and predict the diagnostic class of a sample. Actually, precise diagnosing and classification is crucial for in the treatment of health problem (Mei, Shen, & Ye, 2009).

DOI: 10.4018/IJSI.2019070104
Microarray knowledge analysis are often administrated in keeping with a minimum of 2 totally different and complementary views. In one hand, knowledge cluster (non-supervised classification) aims to spot teams of genes or teams of experimental conditions that exhibit similar expression patterns. In such a context bi-clustering is especially fascinating since it permits the synchronous identification of teams of genes that show similar expression patterns across specific teams of experimental conditions (samples) (Duvaland & Hao, 1999).

The processing of these data, therefore, requires reducing the number of genes to provide a subset of relevant genes and construct a classifier predicting the type of tumor that characterizes a cell sample. The selection of attributes is a complex problem that has already been widely studied, but the dimensions of the biochip data require specific approaches (several thousand genes).

Several machine learning methods have been applied to gene expression data classification. Ben-Dor has proved that Nearest Neighbor Algorithm (NN Algorithm), Boosting Algorithm and clustering-based classification algorithm can all be applied to gene expression data classification (Ben-Dor et al., 2000). Support Vector Machines (SVMs), a classical machine learning algorithm was firstly applied to gene expression data classification by (Grundy et al., 2000). A principal component and discriminant analysis method of tumor classification was proposed by (Xiong, Jin & Li, 2000).

Data mining is one of the important phases of knowledge data discovery, there is a technique which is used to find new, hidden and useful patterns of knowledge from large databases. There are several data mining methods such as Prediction, Clustering, and Classification (Lavanya and Usha Rani, 2011).

Indeed, the extremely rapid growth of the data collected in the databases and the need for effective responsiveness on the part of decision-makers to this new information have stimulated the rapid development of Knowledge Extraction in the last decade. From the Data (ECD), the ECD, or Knowledge Discovery in Databases, is a “non-trivial process of identifying unknown, valid and potentially exploitable structures in databases” [FPSS96]. The ECD consists of four phases (see Figure 1): data acquisition and storage (Data Warehousing), data pre-processing, data mining and post-processing, our work is mainly at the level of stage three: data mining. This step integrates both the choice of the appropriate modeling and the method to be used as well as its application to the search for underlying data structures and to the creation of explanatory and/or predictive models.

This step can be broken down into three major tasks: discrimination (supervised classification), categorization or clustering, and the search for association rules. These tasks can, by their formalization, be modeled in combinatorial optimization problems.

From the formalization of these various tasks of extraction of knowledge to problems of combinatorial optimization, it is necessary to implement methods of resolution. Usually, the methods used in this framework are of three types: exact methods, metaheuristics, and task-specific heuristic methods.

The selection of attributes is a complex problem that has already been widely studied, but the dimensions of the biochip data require specific approaches (several thousand genes).

In this study, we considered two datasets for proposed approach to analyzing the performance of decision tree classifier, a C4.5 algorithm with hybrid feature selection methods to find out whether the same feature selection method may lead to best accuracy and best AUC on various datasets of the

![Figure 1. ECD through four phases](image-url)
Related Content

Formalization of Expert Knowledge About the Usability of Web Pages Based on User Criteria Aggregation

QSE: Service Elicitation with Qualitative Research Procedures
[www.igi-global.com/chapter/qse/117924?camid=4v1a](www.igi-global.com/chapter/qse/117924?camid=4v1a)
An Agent Based Formal Approach for Modeling and Verifying Integrated Intelligent Information Systems
[www.igi-global.com/chapter/agent-based-formal-approach-modeling/23793?camid=4v1a](www.igi-global.com/chapter/agent-based-formal-approach-modeling/23793?camid=4v1a)

High-Level Design Space Exploration of Embedded Systems Using the Model-Driven Engineering and Aspect-Oriented Design Approaches
[www.igi-global.com/chapter/high-level-design-space-exploration/36340?camid=4v1a](www.igi-global.com/chapter/high-level-design-space-exploration/36340?camid=4v1a)