AIN SHAMS UNIVERSITY
Faculty of Computer
& Information Sciences
Abbassia, Cairo, Egypt

Data Mining Techniques in Gene Expressions

A Thesis submitted to the Department of Computer
Science, Faculty of Computer and Information Sciences,
Ain Shams University,

In partial fulfillment of the requirements for the degree of
Master of Computer Science

By

Basma Ali Maher
B.Sc. in Computer Science, 2009
Computer Science Department, Faculty of Computer and Information Sciences,
Zagazig University

Under the Supervision of

Prof. Dr. Abd El-Badeeh M. Salem
Professor of Computer Science,
Computer Science Department,
Faculty of Computer and Information Sciences,
Ain shams University

Prof. Dr. El-Sayed Mohamed El-Horbaty
Professor and head of Computer Science Department,
Faculty of Computer and Information Sciences,
Ain shams University

Dr. Abeer Mahmoud Mahmoud
Lecturer,
Computer Science Department,
Faculty of Computer and Information Sciences,
Ain shams University

2014

Acknowledgement

First and foremost, I would like to thank ***Allah*** for giving me the opportunity and the strength to accomplish this work.

I would like to express my greatest and deepest gratitude to **Prof. Dr. Abd El-Badeeh Mohammed Salem and Prof. Dr. El-Sayed Mohamed El-Horbaty** for planning and full supervision of this work, continuous advises, guidance and greatest help in interpretation of the results. Thanks a lot for their valuable suggestions in writing the thesis.

A very deep and special thanks to my direct supervisor **Dr. Abeer Mahmoud Mahmoud** for her continuous help, time and effort during the research years of this study.

I would like to express my deep grateful thanks to **My Family** who encouraged me too much during the progress of the work.

Abstract

In recent years, the rapid developments in the genetics field have generated a huge amount of biological data. Microarray gene expression data is an important instance of biological data. It has high dimensionality with a small number of samples accompanied with large number of genes. Therefore, using machine learning techniques for knowledge discovery in such data become a rich area for researchers. The mining phase is usually divided into two steps: the gene selection (feature reduction) and the classification process.

Gene selection is a process of finding the genes most strongly related to a particular class. The benefit of this process is to reduce not only dimensionality but also, the danger of presence of irrelevant genes that affect the classification process. Many machine learning approaches are used feature reduction but the study focuses on t-test and class separability. In the other hand, classification is an important data-mining problem that has a wide range of applications. Classification concerns learning that classifies data into the predetermined categories. It is applied to discriminate diseases or to predict outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature. Many machine learning approaches are used classification. In this study, it focuses on Support vector machine and k-nearest neighbor.

Support Vector Machine (SVM) plays a very important role in the data mining classification problem. The structure of SVM depends on kernel functions, where the most commonly used are liner and polynomial. If there are more than two classes in the data set, binary SVMs are not sufficient to

solve the whole problem. To solve multi-class classification problems, the whole problem should be converted into a number of binary classification problems. Usually, there are two approaches. One is the “one against all” scheme and the other is the “one against one” scheme.

On the other hand, K-Nearest Neighbor shows an outstanding performance in many cases of classifying microarray gene expression. For using KNN technique three key elements are essential, (1) a set of data for training, (2) a group of labels for the training data (identifying the class of each data entry) and (3) the value of K for deciding the number of nearest neighbors.

This study proposes a new hybrid reduction approach for the promotion of the cancer classification accuracy that uses two gene selection techniques to confirm the most informative genes and to discard irrelevant genes that affect the classification accuracy. Actually, it applied two machine learning (ML) gene ranking techniques (T-test and Class Separability (CS)) and two ML classifiers; K-nearest neighbor (KNN) and support vector machine (SVM); for exploring and analyzing the process of mining microarray gene expression profiles. In addition, based on these analyses we proposed a hybrid ML reduction approach to enhance the classification accuracy.

It has tested and validated the ML approaches on four public microarray databases; Lymphoma, Leukemia, Small Round Blue Cell Tumors (SRBCT) and Lung Cancer datasets. The experimental results show that the hybrid system achieves enhancement in the classification accuracy better than the SVM and KNN techniques alone. Also, selecting genes from the whole data is better than selecting it from the training data. But excluding the testing samples from the classifier building process, make it more accurately to compare the performance and it make a validation for the system.

List of Publications

[1]. Basma A.Maher, Abeer M. Mohamed, El-Sayed M.El-horbaty and Abd El-Badeeh M.Salem, “Classification of Two Types of Cancer Based on Microarray Data”, Egyptian Computer Science (ECS), Vol. 38, No.2, PP.56-66, 2014.

[2]. Abeer M. Mohamed, Basma A.Maher, “A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data”, Int. Journal of Advanced Research in Artificial Intelligence, Vol. 3, No. 10, PP. 1-10, 2014.

[3]. Abeer M. Mohamed, Basma A.Maher, El-Sayed M.El-Horbaty and Abd El-Badeeh M. Salem, “Applying a Statistical Technique for the Discovery of Differentially Expressed Genes in Microarray Data”, Proceedings of the 4th European Conference of Systems (ECS '13), France, PP. 220-227, 2013.

[4]. Abeer M. Mohamed, Basma A.Maher, El-Sayed M.El-Horbaty and Abd El-Badeeh M. Salem. “Analysis of Machine Learning Techniques for Gene Selection and Classification of Microarray Data”, Proceedings of the 6th Int. Conf. on Information Technology, Cloud Computing, Jordon, 2013.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADCA	Adenocarcinoma
AI	Artificial Intelligence
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ANN	Artificial Neural Network
ANOVA	Analysis-of-Variance
BL	Burkitt Lymphomas
CC	Correlation Coefficient
CLL	Chronic Lymphocytic Lymphoma
CS	Class Separability
CV	Cross Validation
DEGs	Differentially Expressed Genes
DLBCL	Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic Acid
E	Entropy
ED	Euclidean Distance
ELM	Extreme Learning Machine
EWS	Ewing Families of Tumors
FC	Fisher Discriminate Criterion
FL	Follicular Lymphoma
FNN	Fuzzy Neural Network
FS	F(x) Score
KDD	Knowledge Discovery Database
KNN	K-Nearest Neighbor
LDA	Linear Discriminate Analysis
MD	Mean Difference
ML	Machine Learning
MPM	Malignant Pleural Mesothelioma
NB	Neuroblastoma
NPPC	Nonparallel Plane Proximal Classifier
PCA	Principal Component Analysis
RBF	Radial Basis Function
RFE	Recursive Feature Elimination
RMS	Rhabdomyosarcoma

RNA	Ribonucleic Acid
SANN	Subsequent ANN
SNR	Signal to Noise Ratio
SRBCT	Small Round Blue Cell Tumors
SVM	Support Vector Machine
SVM-OAA	Support Vector Machine One Against All
TS	T-test Statistics

Chapter 1

Introduction

1.1 Problem Definition

Creatures consist of organisms and every organism carries the same genetic information. This genetic information is represented in the form of genes, where only a subset of these genes is active or expressed. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science [1]. The broad use of machine learning techniques and their applicability in the different areas of bioinformatics reported a success resolving biological problems because it facilitates the process of analysing such data sets and extracting the important hidden knowledge. The cancer classification based on the microarray data is one example of this type of analysis.

Simply, Microarray gene expression data refers to such repositories of gene information that made the technology of modern biological research. Its goal is to understand the regulatory mechanism that governs protein synthesis and activity of genes. All the cells in an organism carries equal number of genes yet their protein synthesis can be different due to regulation. Protein synthesis is regulated by control mechanisms at different stages [2]:

1) Transcription
2) RNA splicing
3) Translation
4) post transitional modifications

Furthermore, analyzing the gene with respect to whether and to what degree they are expressed can help characterize and understand their functions. It can further be analyzed how the activation level of genes changes under different conditions such as for specific diseases (e.g. cancers are generally caused by abnormalities in the genetic material of the transformed cells or change in their activation or function) [3]. Actually, microarray represents a powerful tool in biomedical discoveries and harnessing the potential of this technology depends on the development of appropriate mining approaches [4][5][6].

Microarray techniques provide a plat form where one can measure the expression levels of thousands of genes in hundreds of different conditions. Actually, there is a high redundancy in microarray data and numerous genes contain inappropriate information for precise classification of diseases or phenotypes [7]. Therefore, the amount of data generated by this technology presents a challenge for the biologists to carry out analysis [8].

The mining phase in the knowledge discovery process can be defined as the process of discovering interesting and unknown patterns from large amounts of data stored in information repositories [9][10]. The mining task could be one of regression, summarization, clustering and classification [9]. In microarray data, classification is momentously necessary for cancer diagnosis and treatment. Specially, in classification analysis of microarray data; where the data has high dimensionally; gene selection is one of the critical aspects, where the objective is reaching an efficient gene selection approach that can drastically ease computational burden of the subsequent classification task, and can yield a much smaller and more compact gene set without the loss of classification accuracy.