Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 300–309

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Selecting significant marker genes from microarray data by filter approach for cancer diagnosis

Sara Haddou Bouazza[1]*, Khalid Auhmani[2], Abdelouhab Zeroual[1], Nezha Hamdi[1]

[1]*Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco*
[2]*Department of Industrial Engineering, National School of Applied Sciences, Cadi Ayyad, Safi, Morocco*

Abstract

In machine learning, feature subset selection phase is the process of selecting a small subset of the most relevant features for use in a model construction. The main goal of this paper is to perform a comparative study between features subset selection methods applied to the DNA microarray dataset and to investigate the strength of each method. The studied methods are: F test, T test, Signal to noise ratio (S/R), ReliefF and Pearson product-moment correlation coefficient (CC). This study is carried out using the dataset of the five cancers; Leukemia, Lung, Lymphoma, Central Nervous System and Ovarian cancers. The Evaluation of the studied methods has been done by using the supervised classifiers: K Nearest Neighbors (KNN), Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC) and Naïve Bayes classifier (NV).
The purpose of classification is to predict the presence of cancer. The classification accuracy is measured for each selected subset of genes. Results show that the combination between S/R selection method and the KNN classifier present the highest accuracies for different cancers; 100% for only 13 relevant genes in Leukemia cancer, 100% for 21 genes in Lung cancer, 100% for 4 genes in Lymphoma cancer, 76.7% for 6 genes in CNS cancer, 100% for 30 genes in ovarian cancer.

Keywords: Biochip data; feature selection; classification; Data mining; image processing

1. Introduction

Biotechnology is the use of living systems and organisms to develop any technological application that uses biological systems. It permits the measure of the information in genes, and offer diagnostic tools of all kinds of

* Corresponding author. Tel.: +212 648 71 94 93; fax: +212 648 719 493.
*E-mail address:*sara.hb.sara@gmail.com

1877-0509
10.1016/j.procs.2018.01.126

cancers.

Available datasets are characterized by a limited number of observations and a high number of features. The datasets need to be reduced into a limited number of relevant features and then build a classifier predicting tumor type of the sample.

In this paper, we will study different selection methods and classifiers, and we will test them on different

datasets. To do this, we will use five databases: Leukemia, Lung, Lymphoma, Central Nervous System, and Ovarian cancers. Afterwards, we select relevant features from the entire original cancer dataset, using five selection methods (F test, T test, Signal to noise ratio, ReliefF and Pearson product-moment correlation coefficient). Finally, we will study the performance of each selection method on each cancer dataset, by using five classifiers (K Nearest Neighbors, Support Vector Machines, Linear Discriminant Analysis, Decision Tree for Classification, and Naïve Bayes classifier).

The rest of the paper is organized as follows. In Section 2, we propose the definition of biochips technology, Feature Selection Methods, and classification steps. In Section 3, we compare the performance of different feature selection and classification methods. In section 4 we discuss the results obtained in section 3. Finally, we present conclusions in Section 5.

2. Materials and methods

2.1. *Biochip technology*

A biochip or more Precisely "Deoxyribonucleic acid microarray (DNA)" is a group of DNA spots attached to a solid surface. Experts extract and measure the expression level of genes from DNA chip in different pathological conditions [1; 2; 3; 4; 5; 6; 7; 8; 9].

The biochip technology is built on the complementary strands of the DNA double helix principle and the hybridization property between two complementary nucleic acid sequences [1; 10; 11; 6].

2.2. Feature Selection Methods

Dimensionality reduction is a necessary step which selects the most relevant and significant features in the original dataset without any transformation while keeping the physical meanings of the original features. It removes redundant features by applying different feature selection methods on the dataset. It's able to improve learning performance, lower computational complexity, build better generalizable models, and decrease the storage space.

There are three common classes of feature selection algorithms: filter, wrapper and embedded methods.

Filter feature selection methods use a statistical measure to assign a score to each feature. Wrapper methods consider the selected subset of features as a search problem. Embedded methods learn which features best contribute to the accuracy of the model while the model is being created.

In this paper, features in every cancer are the genes. We adopt the filter methods which are based on the estimated weight (scores) corresponding to each feature (gene), used to order then to select the most relevant features.

F test

The F test is any statistical test in which the test statistic has an F distribution under the null hypothesis. It gives a score defined as follows [12]:

$$F_{(g)} = \frac{(M1 - M2)^2}{(S1^2 + S2^2)} \quad (1)$$

Where M_k; S_k^2 denotes the mean and standard deviation of the feature (g) for the class k = 1; 2.