Egyptian Informatics Journal xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

Gene expression based cancer classification

Sara Tarek *, Reda Abd Elwahab, Mahmoud Shoman

Faculty of Computers and Information, Department of Information Technology, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 4 September 2016
Revised 23 October 2016
Accepted 6 December 2016
Available online xxxx

Keywords:
Microarrays
Cancer
Classification
Gene expression
Feature selection
Ensemble
K-NN
Bioinformatics
Computer science
Machine learning

ABSTRACT

Cancer classification based on molecular level investigation has gained the interest of researches as it provides a systematic, accurate and objective diagnosis for different cancer types. Several recent researches have been studying the problem of cancer classification using data mining methods, machine learning algorithms and statistical methods to reach an efficient analysis for gene expression profiles.

Studying the characteristics of thousands of genes simultaneously offered a deep insight into cancer classification problem. It introduced an abundant amount of data ready to be explored. It has also been applied in a wide range of applications such as drug discovery, cancer prediction and diagnosis which is a very important issue for cancer treatment. Besides, it helps in understanding the function of genes and the interaction between genes in normal and abnormal conditions. That is done by monitoring the behavior of genes -gene expression data- under different conditions.

In this paper, an effective ensemble approach is proposed. Ensemble classifiers increase not only the performance of the classification, but also the confidence of the results. The motivations beyond using ensemble classifiers are that the results are less dependent on peculiarities of a single training set and because the ensemble system outperforms the performance of the best base classifier in the ensemble.

1. Introduction

Deoxyribonucleic acid or DNA stores genetic information required by all livings in order to build, function and develop. DNA is said to be the blueprint of all living organisms since its components encode all the information needed for maintaining life. This genetic information are preserved and passed from one cell to another during the process of cell division in which a parent cell splits into two new daughter cells. DNA molecules form a double twisted helix bonded together and arranged in a very precise order. The basic four molecular units forming the DNA helix are then sequenced in a particular arrangement such that each component on one strand can only bond with a certain component in the other strand. DNA replicates by breaking the bond between its two strands -the double twisted helix- and each strand form a matching strand, re-bond and re-twist once again.

The genome -the entire DNA sequence- provides a template for the synthesis of a variety of RNA molecules. The main types of RNA are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). One of the major functions of the DNA is to construct proteins which are responsible for carrying out most of cell functions. The process of constructing proteins consists of two major steps; namely, transcription stage in which DNA molecule is transcribed into messenger RNA or mRNA (which is a type of ribonucleic acid RNA); and translation stage, where mRNA is translated into proteins' amino acid sequences to perform cell functions. Once protein is constructed, the gene is said to be *expressed*. The standard technique for measuring gene expression is to measure the mRNA instead of proteins. The reason behind using mRNA sequences is that they hybridize with their complementary RNA or DNA sequences while this property lacks in proteins.

Gene expression level represents the amount of RNA produced in a cell under different biological states. So during cell division process, if the cells suffer from diseases -i.e. cancer or malignant tumors- that cause alteration or mutations in genes, the uncontrollable behavior of the gene will be transmitted to daughter cells. Moreover, certain gene expression values will be affected and hence expression levels can be realized by monitoring the RNA.

The expression levels of thousands of genes can be simultaneously measured under particular experimental environments and conditions due to the significant advancement of DNA microarray technology. This technology made it possible to understand life

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

* Corresponding author.
E-mail addresses: sarah.tarek1@gmail.com (S. Tarek), r.abdelwahab@fci-cu.edu.eg (R. Abd Elwahab), m.essmael@fci-cu.edu.eg (M. Shoman).

http://dx.doi.org/10.1016/j.eij.2016.12.001
1110-8665/

Please cite this article in press as: Tarek S et al. Gene expression based cancer classification. Egyptian Informatics J (2016), http://dx.doi.org/10.1016/j.eij.2016.12.001

on the molecular level, and enables to generate large-scale gene expression data. Besides, it has led to many analytical insights because it produced large amount of gene data ready to be analyzed rapidly and precisely by managing them using several statistical and machine learning processes.

In order to transform the DNA microarray samples from their analog form, which is DNA sequences printed in a high density array on a glass microscopic slide, into a digital form, which is the gene expression data matrix that can be handled and manipulated, several steps should be done.

The standard technique is to transcribe mRNA from two cells and reverse transcribe both of them into cRNA, label them using fluorescent dyes (red if cancerous and green if benign). Both samples are distributed over the whole microarray in order to hybridize with their corresponding cDNA (labeled cDNA try to bind to their complementary cDNA on the microarray in order to form a double stranded molecule in the process called hybridization). Hybridization thus acts like a detector of the presence of a certain gene. The slide is then scanned to obtained numerical intensities of each dye. Then, the substrate can be scanned as an image that can be manipulated by image processing techniques; the intensity of the colors corresponds to the number of mRNA transcribed for each gene. By comparing the intensity of the colors for a gene under two different experimental conditions, gene expression levels can be monitored. For all genes on a single chip, gene expression value is: $\log_2(I_R/I_G)$ where I_R is the intensity of the red dye and I_G is the intensity of the green dye.

DNA microarray technology provided a considerable amount of important data ready to be explored. This massive amount of data suffers from several issues. First, the process of extracting samples under the right conditions is extremely difficult to meet and it involves high level of noise. Second, there are thousands of gene expressions versus few dozens of samples, which require excluding the irrelevant genes before actual classification. The choice of the discriminating genes between different cancer types or classes is a major research field. Third, several tradeoffs have been revealed such as maintaining accuracy rate versus ensuring generalization, controlling complexity versus improving classifier's performance, improving the performance versus memory requirements. Those factors have affected the efficiency of the cancer classification algorithms.

2. Related work

Okun [1] suggested an ensemble system implemented over Colon dataset. Filter feature selection models are used to mitigate the effect of overfitting. Three Different gene selection methods were implemented; namely, Backward Elimination Hilbert-Schmidt Independence Criterion "BAHSIC" [2], Extreme Value Distribution based gene selection "EVD" [3] and Singular Value Decomposition Entropy gene selection "SVDEntropy" [4]. The ensemble consists of five base classifiers, each utilizes K-nearest neighbor "K-NN" with different values of K "either 3 or 5 nearest neighbors". The choice of K-NN classifier was justified as it doesn't require training which is more suitable for usage with the Colon dataset because of the nature of the microarray data (few dozens of samples with high dimensionality).

Because of the small sample size, bolstered resubstitution error estimation "BRE" is used [5]. The bolstered resubstitution estimator is based on the theory that, more confidence should be attributed to points far from the decision boundary than points near it. Bolstered resubstitution error estimators are low variance and generally low bias as well. It is very competitive in comparison with cross validation and bootstrap error estimators especially for small sample problems [6].

Although the results of the as-is system looks appealing; however, it can be further improved by taking some points into consideration. First, due to the sensitivity of the cancer classification domain, more accurate output results are desired such that the error is minimized. In this existing system, the total ensemble error of the as-is ensemble system is 6.5%, which is relatively high. Second, the existing system has been tested against Colon dataset only, while a more comprehensive system that is tested against various cancer datasets is required. Third, in the case of small sample size where substitution is extremely low biased due to over-fitting, bolstering or spreading the error of the misclassified instances is not practically suitable. This is because it contributes in increasing the bias; specially with over-fitting rules classifiers. Fourth, the choice of the method of combining votes of the base classifiers into one predicted decision should consider the accuracy -or the weight- of the ensemble member. This is in order to determine to what extent an ensemble member should participate in the final decision. Also, the size of the dataset on which the decision has been taken should be considered (e.g. Naïve Bayes combination). Moreover, the choice of the feature selection methods that preserve the semantics of the features 'genes' and further select more informative and relevant features (e.g. Markov Blanket) is a more suitable from biological point of view. Currently, some feature selection techniques on microarray data have been proposed [5–7].

3. Proposed system

We propose an ensemble system which is a set of individually trained classifiers whose decisions are combined typically with majority voting, weighted voting or other relatively simple techniques such as stacking or Naïve Bayes combination. Researches show that generally ensemble classifier outperforms the performance of the best member classifier in the squad [8–10].

The proposed system addresses the first three drawbacks of the existing system; namely, enhancing result accuracy, applying the ensemble technique to more cancer types, and mitigating the effect of over-fitting. Block diagram of the proposed system is presented in Fig. 2. The shaded blocks in the diagram indicate a contribution is done in this block. The following points explain the function of each module as well as the modifications done at each one:

- **Gene Expression Dataset:** In this module the dataset that the system will run is defined. Sequence of actions to be carried out on those datasets is defined amongst file reading, loading and connecting to dataset repository. The proposed ensemble system has been applied to Colon dataset as the case in the existing system introduced by [11]. In addition, the proposed system has been modified, tuned and tested against Leukemia and Breast cancer datasets to boost the confidence in applying the proposed system to different cancer types and to emphasize the suitability of the proposed system to be applied in this sensitive domain.
- **Preprocessing Module:** According to the dataset defined in the gene expression dataset module, preprocessing module prepares the dataset to be manipulated. Preparation includes filtering, thresholding, logarithmic transformation and data normalization. Those procedures are essential to be done before actual classification take place.
- **Gene Selection Module:** In terms of cancer classification, this massive number of microarray data doesn't bring more discriminative power; degrade the accuracy of the classifier. Thus, features need to be decreased to a feature subset with the most significant features or genes that are capable of discriminating different classes. So, the objective of feature selection -also

Please cite this article in press as: Tarek S et al. Gene expression based cancer classification. Egyptian Informatics J (2016), http://dx.doi.org/10.1016/j.eij.2016.12.001