Applied Soft Computing 35 (2015) 43–51

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization

V. Elyasigomari, M.S. Mirjafari, H.R.C. Screen, M.H. Shaheed*

School of Engineering and Materials Science, Queen Mary, University of London, London E1 4NS, United Kingdom

ARTICLE INFO

Article history:
Received 8 September 2014
Received in revised form 15 June 2015
Accepted 15 June 2015
Available online 22 June 2015

Keywords:
Cancer classification
Gene selection
Clustering
Evolutionary algorithms
Cuckoo optimization algorithm
COA-GA

ABSTRACT

This research presents an innovative method for cancer identification and type classification using microarray data. The method is based on gene selection with shuffling in association with optimization based unconventional data clustering. A new hybrid optimization algorithm, COA-GA, is developed by synergizing recently invented Cuckoo Optimization Algorithm (COA) with a more traditional genetic algorithm (GA) for data clustering to select the most dominant genes using shuffling. For gene classification, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) artificial neural networks are used. Literature suggests that data clustering using traditional approaches such as K-means, C-means and Hierarchical do not have any impact on classification accuracy. This is also confirmed in this investigation. However, results show that optimization based clustering with shuffling increase the classification accuracy significantly. The proposed algorithm (COA-GA) not only outperforms COA, GA and Particle Swarm optimization (PSO) in achieving better classification performance but also reaches a better global minimum with only few iterations. Higher accuracy is observed to have achieved with SVM classifier compared to MLP in all datasets used.

1. Introduction

Biological systems can be viewed as information management systems with a basic instruction set stored in each cell's DNA as genes. For most genes, their information is enabled when they are transcribed into RNA, which is then translated into the proteins that form much of a cell's machinery; the phenomena which is known as gene expression [1]. The vast majority of fatal diseases have a unique gene expression profile which can be observed using microarray technology [2].

One of the important fields that gene profiling can contribute to is cancer classification and profiling tumours. Since some classes of tumours can be better treated with certain drugs than others, gene expression profiling can allow the development of more appropriate personalized treatment plans for individuals [3]. However, with extensive numbers of genes requiring analysis, it is extremely complex to approach this manually. For this reason, pattern recognition, artificial intelligence and Statistical techniques are applied in DNA microarray research. In general, the classification of cancer using microarray data involves data acquisition and pre-processing, gene selection and classification [4]. Classification performance obtained through these processes is then evaluated.

Data acquisition in microarray experiments is performed in laboratories and the data is stored in the form of gene expression ratio. Data pre-processing in microarray technology is an important initial step before data analysis is carried out [5]. Pre-processing removes systematic errors between arrays introduced by hybridization and scanning which are performed in the laboratories.

After pre-processing, the data can be represented in the form of a matrix, where each row in the matrix (see Fig. 1) corresponds to a particular gene and each column could either correspond to an experimental condition or a specific time point at which expression of the genes has been measured [6]. Microarray data in the form of such a gene expression matrix is publicly available to download through websites such as NCBI and ENCODE.

Gene selection is an important aspect for microarray data analysis. Applying statistical and computational methods to microarray data is a big challenge, as this type of data has a high dimension. This is frequently referred as '*the curse of dimension*' in the literature [7]. In diseases such as cancer, only a few genes are generally informative and require analysis [8]. The aim of gene selection is to select those important genes that contribute to cancer and eliminate the rest of the genes, so that the dimension of the data is reduced for further investigation [9]. There are several

* Corresponding author. Tel.: +44 020 7882 3319.
E-mail address: m.h.shaheed@qmul.ac.uk (M.H. Shaheed).

http://dx.doi.org/10.1016/j.asoc.2015.06.015
1568-4946/

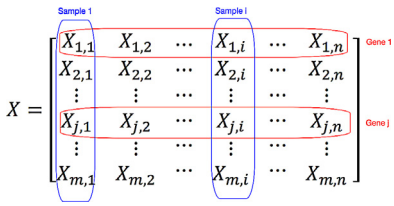

Fig. 1. Gene expression matrix.

disadvantages when the number of genes is significantly more than the number of samples. For instance, both the processing time and the chance of misclassification are increased [10].

Once the genes are selected, the classification procedure follows. In classification, a class predictor, which predicts if the sample is healthy or cancerous, is usually designed, by utilizing the already available data from different diagnostic classes, which are referred to as training or learning samples. In this procedure, first the classifier is trained and then the classifier is used to find the diagnostic class of new samples [11]. There are many approaches that can be used for microarray gene expression classification, like k nearest neighbours (k-NN) [12], Support Vector Machines (SVM) [13], Multilayer perceptron (MLP) [14], or other types of Artificial Neural Networks (ANNs) [11].

It is noted from the literature that data clustering (grouping data) has also been used as a method for cancer classification in a number of investigations [15]. In cluster analysis, there is no prior information on the group structure of the data. Clustering methods divide a set of n genes into g groups so that within group similarities are larger than between group similarities [16]. Clustering techniques can be used in microarray analysis to facilitate visual display and interpretation of experimental results. Last but not the least, clustering suggests the presence of subgroups genes that behave similarly [17].

In this study, the effect of integrating data clustering into the classification procedure is investigated. Three traditional clustering methods known as k-means, fuzzy c-means and hierarchical clustering are utilized, alongside three optimization clustering methods, called Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Cuckoo Optimization Algorithm (COA). Finally, a new hybrid clustering algorithm combining COA and GAs, COA-GA, is proposed. The methods are applied to two microarray datasets and the performance is assessed in terms of accuracy, sensitivity, and specificity. The objective is to investigate whether optimization based clustering could improve the accuracy of cancer classification results compared to situations in which no clustering or traditional clustering methods are used.

2. Microarray data used in this investigation

Microarray data for 4 cancer types (leukaemia, colon, lymphoma and prostate) are considered in this investigation. A gene expression data set for colon cancer (Notterman Carcinoma data) is obtained from (http://genomics-pubs.princeton.edu/oncology/), which is publicly available. The data consists of 7457 genes and each gene has 38 samples in which 18 samples correspond to healthy cases and 18 samples are cancerous. The gene expression profile for leukaemia is obtained from Broad Institute (www.broadinstitute.org/cancer/pub/all_aml) and consists of 7129 genes and 73 samples, where 25 samples correspond to acute myeloid leukaemia (AML) and the remaining 48 samples correspond to acute lymphoblastic leukaemia (ALL). A gene expression data set for Lymphoma cancer is obtained from (http://datam.i2r.a-star.edu.sg/datasets/krbd/), which is publicly available. The data consists of 4026 genes and each gene has 47 samples, 24 of them from "germinal centre B-like" group, while 23 are "activated B-like" group. Finally a gene expression data from prostate is obtained from Broad Institute (http://www.broadinstitute.org/cgi-bin/cancer/publications/view/75), which contains 126,000 genes where each gene has 102 samples in which 52 samples correspond to tumour and 50 samples correspond to healthy cases.

3. Methodology

The general methodology used in this study is illustrated in Fig. 2. First the 25 most informative genes (best genes) are determined using shuffle method. It is a two-stage process involving clustering with optimization and selection of genes. The selected genes are then fed to the classifiers (MLP or SVM). Finally sensitivity, accuracy and specificity are calculated and compared with those of the existing techniques to quantify the performance of the method.

3.1. Gene selection based on shuffle technique

In cancer classification using clustering based gene selection (grouping genes before gene selection), changing the number of clusters, results in selection of different sets of genes (see Eq. (5)). As such, different classification accuracies are acquired. The differences in the selected genes could occur because the initial centroid protocol for clustering is not specified, but selected randomly [18]. Since the clustering outcome is highly dependent on the initial centroids, it often transpires that better results would have been achieved with other initial points. The standard solution is to try the algorithm few times with different initial points [19]. It seems differences in the selected genes can have affect on the classification performance, so by creating a method to decrease probability of changing new genes, the accuracy of the classification performance could increase.

In the case of the shuffle technique, the data is clustered 6 times, setting different numbers of clusters, ranging from 1 to 6, in each case. This means that in the first run the number of clusters is set to one implying no clustering is used, hence the algorithm goes straight to the gene selection step and selects the top 20 genes. In the second run, data is partitioned in two clusters, while the clustering algorithm iterates 100 times to minimize the cost function to achieve more accurate clusters (see Section 3.1.1). After 100 iterations of the clustering algorithm, gene selection is carried out according to the number of clusters and the population in each cluster (see Section 3.1.2). Therefore, depending on the number of clusters different sets of genes are selected. Similar procedure, as in run 2, carries on until the final run in which data is clustered into six partitions.

As a result of 20 genes selected in each run, a total of 120 genes are finally selected from 6 runs, and are scored according to how many times they were repeated in each case. Finally the 25 genes with the highest score are extracted. These are fed into the MLP and SVM classifiers. Finally sensitivity, accuracy and specificity for both SVM and MLP are calculated. Since the selected genes are the outcome of six separate attempts at gene selection, the shuffle technique introduces more robust gene selection. The two stages involved in the shuffle technique, clustering and gene selection, are described in more detail below.

3.1.1. Clustering with optimization

In this study a new hybrid optimization algorithm, COA-GA, is developed merging the recently invented COA [20] and the traditional GA algorithms for data clustering. The results are compared with the GA, the PSO and traditional approaches, such as K-means, C-means and Hierarchical clustering methods. These traditional clustering methods are explained in a number of articles [21–23]. Genetic algorithm is another evolutionary computing method,