Journal of Biomedical Informatics 67 (2017) 11–20

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification

V. Elyasigomari, D.A. Lee, 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 April 2016
Revised 24 January 2017
Accepted 31 January 2017
Available online 3 February 2017

Keywords:
Gene selection
Minimum redundancy and maximum relevance (MRMR)
Evolutionary algorithms
Cuckoo optimization algorithm
Harmony search algorithm
COA-HS

ABSTRACT

For each cancer type, only a few genes are informative. Due to the so-called 'curse of dimensionality' problem, the gene selection task remains a challenge. To overcome this problem, we propose a two-stage gene selection method called MRMR-COA-HS. In the first stage, the minimum redundancy and maximum relevance (MRMR) feature selection is used to select a subset of relevant genes. The selected genes are then fed into a wrapper setup that combines a new algorithm, COA-HS, using the support vector machine as a classifier. The method was applied to four microarray datasets, and the performance was assessed by the leave one out cross-validation method. Comparative performance assessment of the proposed method with other evolutionary algorithms suggested that the proposed algorithm significantly outperforms other methods in selecting a fewer number of genes while maintaining the highest classification accuracy. The functions of the selected genes were further investigated, and it was confirmed that the selected genes are biologically relevant to each cancer type.

1. Introduction

Over the last two decades, the advent of DNA microarray technology has provided opportunities for personalized medicine by analyzing the expression levels of thousands of genes simultaneously [1]. Microarray technology has recently been used to determine subtypes of certain cancers based on differences in the expression levels of key genes [2–4]. This approach provides detailed information on the genetic makeup of any individual cancer patient, thereby potentially improving the accuracy of treatment decisions made by physicians [5].

During microarray analysis, the number of genes is significantly higher than the number of samples [6,7] and classification to a high level of accuracy is challenging due to the phenomenon of dimensionality [8,9]. To overcome these problems, gene selection mechanisms have been introduced in which only the most important genes are selected and used for classification purposes [10–13]. There are several advantages to this process of minimizing the number of genes and selecting only meaningful genes that are more predictive during classification. By having fewer genes, not only is the processing time for classification significantly decreased, but the chance of misclassification is also reduced. Furthermore, using a large number of genes as input into the classifier can cause the classifier to be over-fitted [14].

Gene selection methods can be categorized into three main approaches based on their interaction with the classifier, namely, filter methods, wrapper methods and embedded methods [14,15]. Filter methods assess the relevance of genes by examining only the general characteristics of the data and ignoring the impact of selected genes on the classification performance [16]. Wrapper gene selection initiates a search procedure in the space of possible gene subsets. The selected genes are then evaluated based on their power to improve classification accuracy [17–19]. In the embedded gene selection method, feature selection is linked to the classification stage; however, this connection is much stronger than in the wrapper method. This is because gene selection in embedded methods is included in the classifier construction and the classifier is used to provide a criterion for feature selection [20,21] (see Fig. 1). More recently, evolutionary algorithms developed for gene selection have been utilized within the framework of wrapper methods [22,23]. Finally, a new gene selection approach by means of shuffling based on data clustering was proposed, suggesting that optimization-based clustering could select more informative genes to enhance classification accuracy [24].

Each gene selection approach has advantages and disadvantages [14]. For example, although the filter method is simple and computationally efficient, its performance lags behind other

* Corresponding author.
E-mail address: m.h.shaheed@qmul.ac.uk (M.H. Shaheed).

http://dx.doi.org/10.1016/j.jbi.2017.01.016
1532-0464/

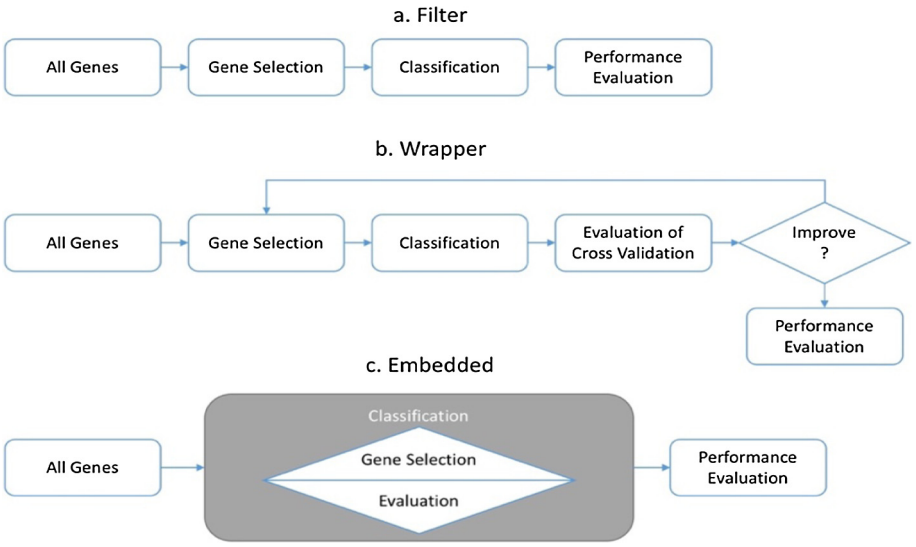

Fig. 1. Featured selection methods.

approaches, since the classifier performs independently and is not involved in gene selection [25]. Conversely, the wrapper and embedded methods, which incorporate the gene selection task into the classification task, can achieve higher classification accuracy but suffer from scalability problems due to their high computational costs and are not practical for large datasets [26,27].

High classification accuracy is, of course, of the utmost importance for personalized medicine. However, biomarker identification is also an area of ongoing research, where it is important to identify a small number of genes to spot patterns (e.g., choosing few genes that are all differentially expressed across different samples) [28,29]. Therefore, in this study, the main objectives were to select the optimum number of the most informative genes that can best distinguish between two cancer types. The gene selection process was performed in two stages. Minimum redundancy and maximum relevance (MRMR) feature selection [30] was first used to select a subset of the most relevant and least redundant genes. The selected genes were then fed into a wrapper setup that combines the proposed COA-HS optimization algorithm with a support vector machine (SVM) as a classifier. The SVM was selected as the classifier in this work, as its classification power has been proven and established by a number of comparative assessments with other algorithms [31–33]. Two-stage gene selection combines the advantages of both the filter and wrapper methods of gene selection. The methods were applied to four microarray datasets and the performance was assessed by the leave one out cross-validation (LOOCV) method.

2. Microarray data

Microarray data for four cancer types (leukemia, prostate, lymphoma, and colon) were used in this study. Gene expression data for leukemia [1] and prostate cancer [34] were obtained from the Broad Institute (www.broadinstitute.org). Gene expression data for lymphoma [35] were obtained from the Lymphoma/Leukemia Molecular Profiling Project (llmpp.nih.gov). A gene expression dataset for colon cancer [36] was obtained from the Princeton University Gene Expression Project (http://genomics-pubs.princeton.edu). Basic information related to the datasets used in this study, including the number of genes, samples and the two classes for each dataset, is provided in Table 1.

3. Methodology

The general methodology is illustrated in Fig. 2. The data were discretized into nine states. After this pre-processing stage, the top 100 genes were selected using MRMR. The selected genes were fed into a wrapper setup consisting of the COA-HS algorithm and the SVM classifier to choose the minimum number of genes that provides 100% accuracy. Finally, the classification performance of the selected genes was measured in terms of accuracy via the LOOCV method. To validate the performance of the COA-HS, the results were compared to those established from other evolutionary algorithms, such as the genetic algorithm (GA), the particle swarm optimization (PSO) algorithm, the harmony search (HS) algorithm, and the cuckoo optimization algorithm (COA). The codes required in this study were written using Matlab 2014a.

3.1. Discretization

The gene expression data were first discretized to reduce the noise and to enhance the accuracy of the classification results [37]. The expression value of each gene was categorized into a nine-state variable based on the mean value (μ) and standard deviation (σ) for that gene. For each gene, the nine states revealed whether the gene was not expressed (state zero) or expressed and how much it was over-expressed (states +1 to +4) or under-expressed (states −1 to −4). Table 2 details the different states utilized in the data discretization.

3.2. First-stage gene selection using minimum redundancy maximum relevance (MRMR)

Gene expression data are typically available in a matrix format (see Fig. 3), where each row represents a gene and each column represents a sample. The last row generally represents the class label for each sample. The class label ($\boldsymbol{C_i}$) for a two-class classification task is defined by either 1 or 2.

The goal of feature selection in a classification task is to identify a subset of features that best characterize the statistical significance of the classification task [38]. MRMR, a filter method of gene selection, identifies those genes that provide more information with respect to the class label of the samples. In this research, mutual information was used for MRMR to determine the relevancy and redundancy of genes and target classes. In this context,