International Journal of Healthcare Information Systems and Informatics
Volume 14 • Issue 1 • January-March 2019

Coupling Multivariate Adaptive Regression Spline (MARS) and Random Forest (RF): A Hybrid Feature Selection Method in Action

Arpita Nagpal, The NorthCap University, Gurugram, India
Vijendra Singh, The NorthCap University, Gurugram, India

ABSTRACT

In this article, a new algorithm to select the relevant features is proposed for handling microarray data with the specific aim of increasing classification accuracy. In particular, the optimal genes are extracted using filter and wrapper feature selection algorithms. Here, the use of non-parametric regression algorithm called Multivariate Adaptive Regression Spline (MARS) followed by proposed Random Forest Statistical Test (RFST) algorithm are being studied. The study evaluates the comparative performance of the results of RFST and MARS with existing algorithms on ten standard microarray datasets. For performance analysis, three parameters are taken into consideration, namely, the number of selected features, runtime, and classification accuracy. Experimental results indicate that different feature selection algorithms yield different candidate gene subset; therefore, a Hybrid approach is applied to determine the best candidate genes to provide maximum information about the disease. The findings foretell that the RFST is performing better on six out of ten datasets whereas MARS is performing better on other datasets.

KEYWORDS

Feature selection, High Dimensional data, Microarray data, Multivariate Adaptive Regression Spline, Random forest

1. INTRODUCTION

Profiling the gene expression of cells to aid in the diagnosis of disease begins logically with the process of preselecting the most informative genes for classification. Even so, one of the greatest challenge when choosing among the genes is that the number of samples collected are often very small as compared to the number of genes. Admittedly, microarray gene expression data suffers from the curse of high dimensionality. The preselection process of identifying informative genes aims chiefly to remove the irrelevant and redundant genes. To date, many gene selection methods have been developed, all of which use different mathematical criteria to filter out and/or extract the subset of relevant genes (e.g., Alshamlan et al., 2015; Bolón-Canedo et al., 2014; Liu et al., 2010; Yeh, 2008; Sardana et al., 2015). A range of different evaluation criteria exists for assessing among

DOI: 10.4018/IJHISI.2019010101

the different feature selection approaches. As per these criteria, feature selection approaches may be broadly classified into four categories as discussed in the next section: filter approach, wrapper approach, embedded and hybrid approach.

Importantly, the application of the MARS algorithm's relative importance score is relatively new in the field of feature selection for cancer microarray data. Hence, we will attempt in this paper to bring in the use of the variable importance score found when applying the Multivariate Adaptive Regression Spline (MARS) algorithm for gene selection. Another key issue to be addressed in the paper is the determination of the optimal and stable candidate gene, which can predict the disease with maximum accuracy. Selecting just a few features may not always be the best approach, as the selected subset of genes may not represent the entire dataset. In addition to using MARS algorithm, this paper will also highlight an algorithm called Random Forest Statistical Technique (RFST) by taking advantage of the variable importance measure returned as an output from the random forest. This importance value has been used to assign weightage to each gene. In RFST, a new statistical test has been proposed to create clusters of similar genes. One representative from each cluster is selected so as to get the final reduced subset of genes.

As well, we will compare RFST and MARS with existing models of fast correlation-based filter or, FCBF (Yu & Liu, 2003) and a fast clustering-based feature subset selection algorithm or, FAST (Song et al., 2013) in terms of the number of gene selected, runtime and classification accuracy. After implementing RFST, MARS, FCBF and FAST filter feature selection algorithms on all collected datasets, it was found that different algorithms select different candidate gene subset. To choose the best candidate genes, which can predict the disease, while returning a high accuracy, a Hybrid feature selection approach has been applied on all the four subsets of genes found using RFST, MARS, FCBF and FAST filter selection algorithms. Classification accuracy is the criterion used in the wrapper feature selection. The subset that maximizes the accuracy is selected as the optimal one. In this work, the approach has been applied on ten binary and multiclass microarray datasets.

The structure of the rest of this paper is as follows. Section 2 overviews the preprocessing filter feature selection approaches and highlights the fundamental principle and preliminaries used to find gene subset in the proposed algorithm. Section 3 discusses the complete methodology while section 4 details the empirical study set up and datasets. Section 5 reports the experimental results found on the publicly available cancer microarray datasets. Finally, our conclusion and an outlook at future work is given in section 6.

2. PRELIMINARIES

Following an overview of the various feature selection approaches, this section introduces the fundamental principle and definitions used in the gene reduction algorithms. It also discusses the basis of MARS algorithm along with the proposed algorithm Random Forest Statistical Test (RFST). RFST is based on the concepts derived from RF algorithm.

2.1. Feature Selection Methods

In a typical filter feature selection approach, feature subset is selected as a preprocessing step before applying any learning and/or classification process. These preliminary processes then become independent of the learning algorithm to be applied. Currently, FCBF and FAST are two major alternative filter feature selection algorithms. These approaches use information theory concept to segregate features; that is, highly correlated features with the class are segregated based on symmetric uncertainty. However, the main disadvantage of FCBF/FAST is the necessity to set a threshold value *a priori*. This has the risk of choosing too many or too few features. Moreover, FCBF considers pairwise correlation among the features rather than joint correlation.