A review of a hybrid feature selection method for DNA microarray data

Laura Rodríguez-Navas[0000−0003−4929−1219]

Universidad Internacional Menéndez Pelayo (UIMP), Madrid, Spain
rodrigueznavas@posgrado.uimp.es

Abstract. The emergence of DNA Microarray technology has enabled researchers to analyse the expression level of thousands of genes simultaneously. The Microarray data analysis is the process of finding the most informative genes as well as remove redundant and irrelevant genes. One of the most important applications of Microarray data analysis is cancer classification. Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. In cancer classification, available training data sets are generally of a small sample size compared to the number of genes involved. Along with training data limitations, this constitutes a challenge to certain classification methods. However, the curse of dimensionality and the curse of sparsity make classifying gene expression profiles a challenging task. One of the most effective methods to overcome these challenges is feature (gene) selection. The feature (gene) selection can be used to successfully extract those genes that directly influence classification accuracy, and to eliminate genes that have no influence on it. This significantly improves calculation performance and classification accuracy. In this paper, I aim to review the correlation-based feature selection (CFS), and the Taguchi-genetic algorithm (TGA) merged into a new hybrid method since the classification accuracy obtained by the proposed method is higher when it is compared to other classification methods from the literature.

Keywords: Feature selection · Taguchi-genetic algorithm · Leave-one-out cross-validation.

1 Introduction

DNA Microarray technology is a powerful tool that helps researchers to monitor the gene expression level in an organism. Microarray data analysis provides valuable results which contribute towards solving gene expression profile problems. One of the most important applications of Microarray data analysis is cancer classification. Cancer may be a genetic disease; the analysis of cancer pathology in the analysis of genes that cause cancer, i.e. the gene whose mutation is responsible for cancer. This reflects the changes in the expression level of various genes. However, classifying the gene expression profile is a challenging task and considered as (NP)-Hard problem [1]. Hence, not all genes contribute to the

presence of cancer. A vast number of genes are irrelevant or insignificant to clinical diagnosis. Therefore, incorrect diagnoses can be reached when all the genes are used in Microarray gene expression classification. There are two main issues related to the analysis of Microarray data; first, the dataset in Microarray is high-dimensional which means it contains several thousand genes (features) and it has low data sparsity, meaning it has a low number of samples, usually tens of samples. Second, gene expression data has a high complexity; genes are directly or indirectly correlated to each other. Standard machine learning methods did not perform well because these methods are best suited when there are more samples than features.

To overcome these issues, dimension reduction or feature (gene) selection algorithms have been applied. Generally, gene selection methods are categorized into three categories: filter, wrapper and embedded methods. The filter approach separates data before the actual classification process takes place and then calculates feature weight values, and thus features that accurately present the original data set can be identified. However, a filter approach does not account for interactions amongst the features. The method in the filter approach category is described in section 2.1. Wrapper models, on the other hand, generally are focused on improving classification accuracy of pattern classification problems and typically perform better (i.e., reach higher classification accuracy) than filter models. However, wrapper approaches are more computationally expensive than filter methods. Several methods in this category have previously been used to perform feature selection of training and testing data, such as genetic algorithm (GA) which is described in section 2.2. And the embedded techniques use an inductive algorithm. The inductive algorithm itself represents the feature selector and the classifier, searching for an optimal subset of features that are built into the classifier. The advantage of embedded algorithms is that they take the interaction with the classifier into account. A disadvantage of embedded algorithms is that they are generally based on a greedy mechanism, i.e., they only use top-ranked attributes to perform sample classification [2][3].

Recently, hybrid and ensemble methods were added to the general framework of feature selection. A hybrid approach is built to take advantage of both filter and wrapper approaches. Thus, it combines the computational efficiency of the filter approach with the high performance of the wrapper approach. A hybrid feature selection approach consisting of two stages is presented in this review. The first stage involves a filter approach that is used to calculate correlation-based feature weights for each feature, thus identifying relevant features. And the second stage constitutes a wrapper approach, i.e. the previously identified relevant feature subsets are tested by a Taguchi-genetic algorithm (TGA), which tries to determine optimal feature subsets. The optimal performance of the hybrid approach is dependent on two factors; the classification accuracy and the number of selected genes.

2 Feature (gene) selection methods

2.1 Correlation-based feature selection

Correlation-based feature selection (CFS) was developed by Hall in 1999 [4]. CFS is a simple filter method that ranks features subsets, based on the correlation between the heuristic evaluation equation (1). CFS is used to select the best combination of attribute subsets via score values from the original data sets and the heuristic evaluation equation is employed to identify the best combination.

The aim of CFS is to reduce the amount of feature to feature correlations while increasing the feature to class correlations. In this paper, they used Weka [5] to implement CFS and used the selected gene subsets to identify different cancer types and various diseases.

$$Merit_s = \frac{k\bar{\gamma}_{cf}}{\sqrt{k + k(k-1)\bar{\gamma}_{ff}}} \tag{1}$$

where $Merit_s$ is the merit of feature subset S containing k features, $\bar{\gamma}_{cf}$ is the average feature and class correlation, and $\bar{\gamma}_{ff}$ is the average feature-feature intercorrelation ($f \in S$).

2.2 Genetic algorithm

2.3 Taguchi method

3 CFS–TGA method

4 Conclusions

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

References

1. Narendra, P.M., Fukunaga, K.: A branch and bound algorithm for feature subset selection. IEEE Computer Architecture Letters 26(09), 917–922 (1977)
2. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. bioinformatics 23(19), 2507–2517 (2007)
3. Yang, P., Zhou, B.B., Zhang, Z., Zomaya, A.Y.: A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. BMC bioinformatics 11(1), 1–12 (2010)
4. Hall, M.A.: Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato (1998)
5. Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H.: Data mining in bioinformatics using weka. Bioinformatics 20(15), 2479–2481 (2004)

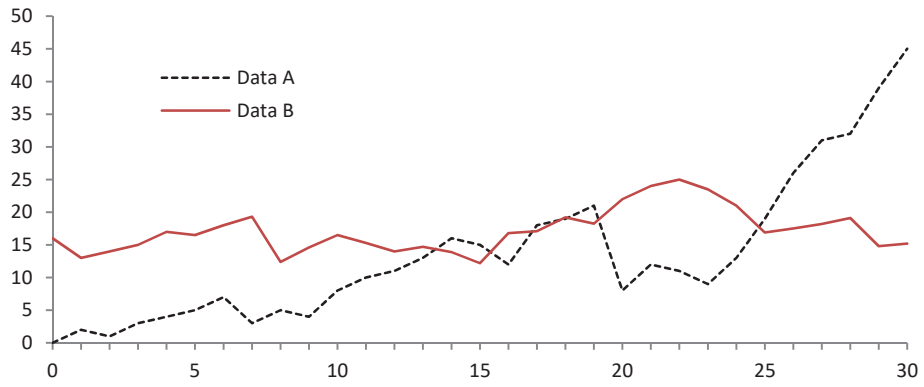

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.