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) that are merged into a new hybrid method, since the classification accuracy obtained by the reviewed hybrid method is higher, 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 mutation that is responsible for cancer and 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 cancer

diagnosis. Therefore, incorrect diagnoses can be reached when all the genes are used in the Microarray gene expression classification. There are two main issues related to the analysis of the Microarray data; first, the dataset in the 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 do 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. 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 are 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 paper. 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, that is 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. This paper use Weka [5] to implement the CFS and use the selected gene subsets to identify different cancer types and various diseases.

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

This is the heuristic evaluation equation 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

A genetic algorithm (GA) was first developed by Holland in 1970. A GA is a stochastic search algorithm modelled on the process of natural selection underlying biological evolution. GA has been successfully applied to many searches, optimization, and machine learning [6].

The main principle of the GA is to generate a population randomly while producing offspring with the same inherited characteristics. The algorithm has evolved into three operations: selection, crossover, and mutation. The selection operation chooses the fittest chromosomes, before allowing them to pass to the next generation. Within the crossover operation, two individuals are then selected via the selection operation process. For each individual crossover then, the operation will select a random crossover point and so the two individuals will swap to produce new offspring. And the mutations are necessary, for there to be a level of diversity maintained in the population.

In this paper, it is proposed a Taguchi method described in section 2.3 and their optimization (described in section 3) that is based on the GA feature selection algorithm, just described.

2.3 Taguchi-Genetic Algorithm

The Taguchi-Genetic Algorithm (TGA) is a hybridization of the Taguchi method and Genetic algorithm (GA) where the Taguchi method is inserted in the crossover and mutation operations. The TGA is normally used as a local search algorithm to select genes for crossover operations.

The Taguchi method was developed by Genichi Taguchi and it is a statistical method of robust design where the processes or products can be analysed and improved by altering relevant design factors. The commonly used Taguchi method [7–9] provides two tools, an orthogonal array (OA) and a signal-to-noise ratio (SNR) for analysis and improvement. The OA used in this paper is a matrix arranged in rows and columns, with each column it is indicated a specific design parameter and each row represent an experimental trial with a combination of different levels for all design factors. This matrix provides a comprehensive analysis of interactions among a balanced set of experimentation runs and systematic comparisons of the different levels of each designed factors. In this paper, the SNR is used to determine the robustness of the levels of each design parameter and it is utilized to analyse and optimize the design parameters for the targets. Conceptually, SNR is represented as (2) and a high SNR level by specifying design parameters represents good results for a particular target.

$$SNR = -10 \log \left(\frac{1}{n} \sum_{t=1}^{n} \frac{1}{y_t^2} \right) \tag{2}$$

3 CFS–TGA method

Chuang et al. [10] proposed a new hybrid method for gene selection that combined Correlation-based Feature Selection (CFS) and the Taguchi-Genetic Algorithm (TGA). This method is the method proposed hybrid method in the paper and it is carried out in two stages. The first stage employs the correlation-based feature selection CFS filter method to remove irrelevant features. And the second stage involves a wrapper approach, employing the TGA methodology to the features that are generated from the filter stage and in doing so identify the best feature subset. Fig. 1 details the CFS–TGA method individual steps of the two stages.

The K-nearest neighbour (KNN) classifier [11, 12] with Leave–One–Out Cross Validation (LOOCV) method [13, 14] is used in the paper to evaluate the proposed hybrid method in terms of classification accuracy. Eleven cancer datasets with binary and multi-class are evaluated. Then, the result of the proposed hybrid method is compared with the result obtained by KNN, measured by the LOOCV method. The comparison shows that the proposed hybrid method achieves the highest classification accuracies in ten datasets where six datasets achieves an accuracy of 100%.

Considering the result, it obvious that the hybrid method shows superior performance in terms of high accuracy and the small number of selected genes. This is because the hybrid algorithm deal perfectly with high dimensionality and overfitting problems by applying the filter approach first as a pre-processing step in order to reduce the dimensionality of the Microarray gene expression profiles.

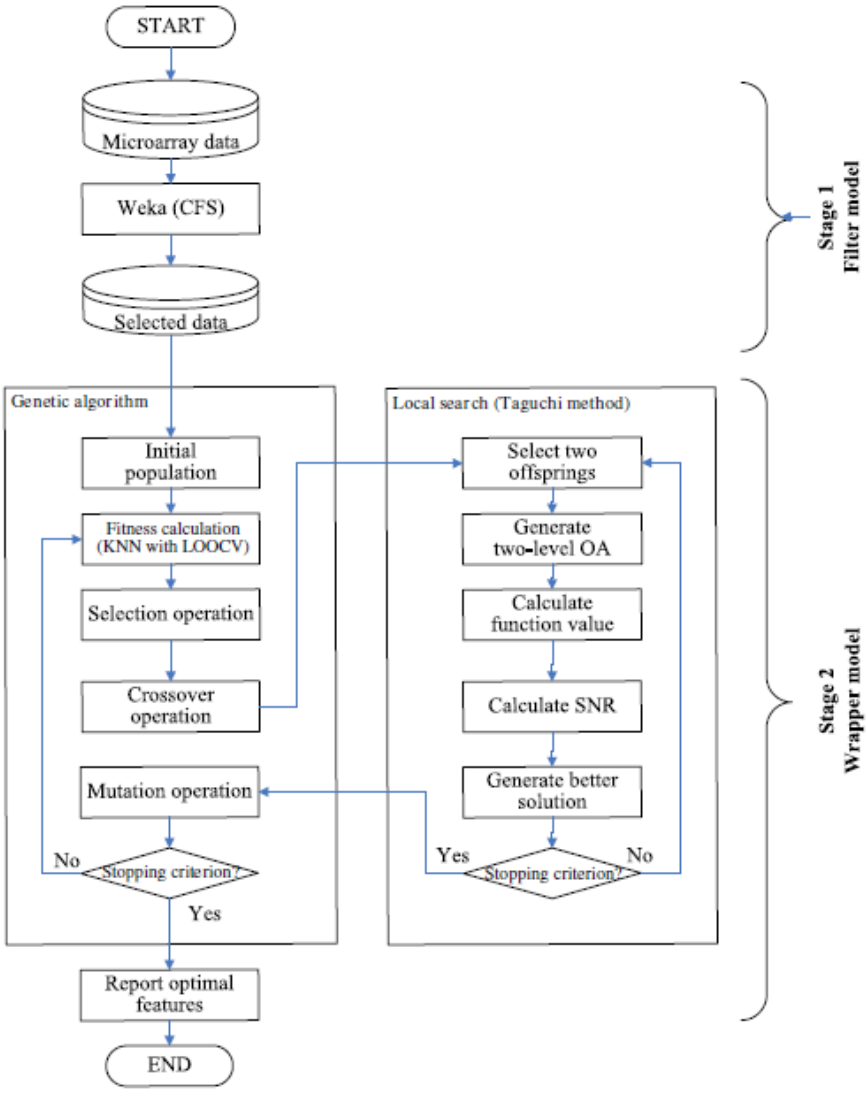

Fig. 1. CFS–TGA method individual steps of two stages.

4 Conclusions

How it is commented in the introduction, the Microarray cancer classification suffers from an overfitting problem. This is due to the curse of dimensionality: a small number of sample and many features. Thus, to avoid it I highly recommend applying the CFS–TGA method in order to adopt filtering before starting the classification process.

Moreover, in this type of problem, the parameters fitting is not an easy process. Because it affects classification accuracy. In cancer gene expression data classification, the parameter fitting is depending on the gene expression datasets and the applied feature selection and classification and methods. Thus, the different dataset has different parameters value, which is not fit for all algorithms.

Microarray data analysis provides valuable results which contribute towards solving gene expression profile problems. As I mentioned before one of the most important applications of Microarray data analysis is cancer classification. Classification is challenging due to the high dimensionality found in a small sample size of gene expression data. The most practical method to overcome these challenges is therefore a feature selection technique that employs hybrid algorithms. In order to review and compare a hybrid algorithm with other types of algorithms from literature, I have chosen this paper. And I can conclude that among all other applied algorithms from the literature the CFS–TGA method achieves the highest accuracy with relatively small numbers of selected genes.

Therefore, as future work, I recommend applying this hybrid gene selection algorithm based on a filter and wrapper approach to identify the most informative genes for cancer classification.

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)
6. Holland, J.H.: Adaptation in natural and artificial systems, university of michigan press. Ann arbor, MI 1(97), 5 (1975)
7. Tsai, J.T., Liu, T.K., Chou, J.H.: Hybrid taguchi-genetic algorithm for global numerical optimization. IEEE Transactions on evolutionary computation 8(4), 365–377 (2004)
8. Taguchi, G., Chowdhury, S.: Robust engineering: learn how to boost quality while reducing costs & time to market. McGraw Hill Professional (1999)

9. Wu, Y., Wu, A.: Taguchi methods for robust design. Amer Society of Mechanical (2000)
10. Chuang, L.Y., Yang, C.H., Wu, K.C., Yang, C.H.: A hybrid feature selection method for dna microarray data. Computers in biology and medicine 41(4), 228–237 (2011)
11. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory 13(1), 21–27 (1967)
12. Fix, E.: Discriminatory analysis: nonparametric discrimination, consistency properties, vol. 1. USAF school of Aviation Medicine (1985)
13. Cawley, G.C., Talbot, N.L.: Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition 36(11), 2585–2592 (2003)
14. Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36(2), 111–133 (1974)