The Performance Review of mRMR for Gene Selection and Classification of DNA Microarrays

IOP Conf. Series: Materials Science and Engineering **551** (2019) 012040 doi:10.1088/1757-899X/551/1/012040

The Performance Review of mRMR for Gene Selection and Classification of DNA Microarrays

Norfadzlan Yusup[1] and Azlan Mohd Zain[2]

[1]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Sarawak, Malaysia
[2]Applied Industrial Analytics Research Group (ALIAS), School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310 Skudai, Johor Darul Takzim, Malaysia

E-mail: fadzlanyusuf@gmail.com

Abstract. There are two main stages in the classification of DNA microarray data. The first stage is known as gene selection and the second stage is the classification of selected genes. The number of genes produced in high-dimensional microarrays is enormous, and only some of these genes help to identify a particular disease. The selection of relevant or informative genes that provide sufficient information about the condition is therefore essential. Gene selection is vital in reducing the data dimensionality which can ease the workload of the computer and increase the high classification performance. In this paper, we review the recent performance on one of the most popular filter based gene selection technique, maximum relevance minimum redundancy (mRMR). We also discuss several current improvements on the mRMR method.

1. Introduction

For Deoxyribonucleic Acid (DNA) microarrays are extensively used in the medical field and considered vital because they enable researchers to 1) identify what caused the diseases 2) how the diseases are classified and 3) how the diseases are treated. The classification is an essential problem in the DNA microarray [1] and currently has received the most attention in the context of cancer research. Gene selection is the process of choosing the most relevant genes in a microarray dataset, and it is an essential task for classification and dimensionality reduction. If the datasets contain irrelevant genes, it will affect not only the training of the classification process but also the accuracy of the model.

Functional classification accuracy is achieved when the model correctly predicted the class labels. Generally, there are three categories of gene selection methods which are Filter, Wrapper and Embedded. Currently, there are two new methods for gene selection; Hybrid and Ensemble method. Classification is a data mining technique that extracts models(classifier) describing significant data classes classifiers where it predicts categorical class labels [2]. There are many types of classifiers used for classification of DNA Microarrays such as Support Vector Machine (SVM), Naïve Bayesian (NB), Random Forest (RF) and Artificial Neural Network (ANN).

2. Filter-based Gene Selection

In the filter method, the gene selection process is independent of the learning process. Gene selection using filter method tend to select redundant features because it did not consider the interactions between features. Once the best features are selected, it will be ranked and evaluated by using either univariate (e.g. Relief F) or multivariate (e.g. mRMR) filter method. Filter method did not necessarily use with classifiers; therefore, it usually used as a pre-processing step [3]. Filter method is computationally less complicated and faster than the wrapper method.

Filter method such as mRMR is employed to generate a subset of relevant gene. mRMR was initially proposed by [4] to reduce the number of genes selected in microarray data for classification. mRMR was also introduced to solve the redundancy issues in ranking approach where the features could correlate among themselves. The fundamental problem with redundancy is that the feature set is not a comprehensive representation of the characteristics of the target phenotypes [4][5]. In mRMR, features that are different to each other are maximized (e.g. mutual Euclidean distances) or their pair-wise correlations are minimized to expand the symbolic power of the feature set. These minimum redundancy criteria are supplement by the usual maximum relevance criteria such as maximal mutual information with the target phenotypes[6].

3. Gene Selection using mRMR

In classifying DNA microarray data, [7] defined the mRMR process as follows. The mRMR method used two mutual information (MI) operations: one between cancer classes and each gene to measure the relevancy, while the second mutual information between every two genes to calculate the redundancy. Therefore, S denotes the selected genes and Rl measures the relevancy of a group of selected genes S that can be defined as follows:

$$\mathrm{Rl} = \frac{1}{|S|}\sum_{G_x \epsilon S} I\left(G_x, C\right), \tag{1}$$

Where $I(Gx\ ,\ C)$ represents the value of mutual information between an individual gene Gx that belongs to S and the cancer class $C = \{c1, c2\ \}$, where $c1$ and $c2$ denote the normal and tumor classes. When the selected genes have the maximum relevance Rl value, it is possible to have a high dependency (i.e., redundancy) between these genes. Hence, the redundancy Rd of a group of selected genes S is defined as,

$$\mathrm{Rd} = \frac{1}{|S|^2}\sum_{G_x, G_y \epsilon S} I\left(G_x, G_y\right), \tag{2}$$

Where $I(Gx, Gy)$ is the mutual information between the xth and yth genes that measures the mutual dependency of these two genes.

4. Classification Performance and improvement on mRMR technique

In this section, we review some of the recent classification performance of DNA microarrays data using mRMR as a gene selection technique. In [7], mRMR was employed to select the most informative genes in six DNA microarray datasets. In this research, mRMR was used to predict top relevant genes that give 100% accuracy. In [8], mRMR with SVM and kNN classifiers were used to choose n top genes of the relevant genes in microarray data. From the result, the number of selected genes by mRMR-SVM is smaller than those of mRMR-kNN with more than 70% classification accuracy achieved in four datasets. In the research by [9] and [10] also achieved excellent results of

mRMR classification using SVM with an average accuracy of 93.06% and 84.15% for all datasets with 50 top-ranked selected genes. In [11] the performance of mRMR experimented on six DNA microarray datasets. Based on their experiments, mRMR with NB classifier performs better than 1NN classifier concerning classification accuracy and the number of selected genes. The performance of mRMR is also better than Relief F and Fast Correlation-Based Filter (FCBF) concerning classification accuracy and the number of selected genes.

The authors in [12] review the performance of four feature ranking (IG, Relief-F, mRMR, SVM-RFE) based on top 10 and top 50 features in nine DNA microarray datasets. The authors also review another three standard feature selection technique (CFS, FCBS and INTERACT). Using mRMR as feature selection, the results of Distribution optimally balanced stratified cross-validation (DOB-SCV) with SVM classifier is better than C4.5 and NB particularly for Colon, Gli85 and Ovarian. The authors in [13] analyze the feature selection techniques of Random Forest Feature Selection (RFFS), Random Forest Feature Selection-Grid Search (RFFS-GS), and mRMR algorithm for comparative experiments in two gene expression datasets. Regarding classification accuracy and Area Under Curve (AUC), mRMR achieved the best results in breast cancer datasets. In [14] the authors compare the performance of four feature selection techniques of mRMR, Chi-Square, Relief F, Effective Range Based Feature Selection (ERGS) with their proposed inter feature effective range overlap technique with top 10 to 60 features selected using NB classifier. With top 10 selected number of features, the results of mRMR are better than other techniques mainly in 11_Tumors, Brain Tumor 1 and Prostate Tumor datasets. The mRMR was also employed in [15] where it identifies the top relevant gene that gives 100% classification accuracy using SVM classifier. mRMR is an advanced processing (filter) stage in which a new subset of data is subsequently transmitted to the wrapper algorithm (e.g. Genetic Bee Colony (GBC).

There are a few improvements that have been proposed by some researchers to improve the performance of mRMR further. Although mRMR is a fast and greedy heuristic, it does not guarantee to find a globally optimal solution. In [16] mRMR was extended by adding a greedy search for mRMR. The experiments used five DNA microarray datasets, and the performance was evaluated based on selection time for 200 features. The proposed approach outperforms the mRMR drastically concerning computation time across three different platforms (CPU, GPU and Parallel Computing). The mRMRe parallelized was applied in [17] with Random Forest classifier to classify Cancer Genome Project (CGP) and Leukaemia microarray data. From the research, the reported selected number of genes is 10, and the out of bag (OOB) error was 0.06 and 0.02 for CGO and Leukemia data respectively. From the literature, the number of the selected gene varies among the researchers. The researchers specified no standard number of selected top *n* genes. For *n* number of the top-ranked gene is usually chosen between 10-50, 50-100, and more than 100 genes. A high number of genes chosen usually gives high classification accuracy [7] [18] [15]. In the research [7] [9], it was reported the standard number of the top-ranked gene to be selected is 50. Researchers mostly prefer the classification of using SVM classifier due to its stability, and it shows good results with mRMR compared to another classifier such as NB and *k*N[8]. From our review, the mRMR method is usually employed as a filter in the hybrid filter-wrapper classification because using mRMR single-handedly does not guarantee to find a globally optimal solution. In hybrid filter-wrapper approach, mRMR was used to reduce the number of the feature set. Then, this reduced feature set will be fed to the next process using various metaheuristics algorithm [19][20].

5. Conclusions

In the literature, mRMR is one of the most popular multivariate filter technique as gene selection in classifying DNA microarrays. The advantage of the mRMR filter method is it can reduce the redundancy in the feature set. However, mRMR does not guarantee to find a globally optimal solution. Therefore mRMR is usually applied as a pre-processing (filter) process in the hybrid filter-wrapper feature selection model. mRMR was employed to reduce the number of irrelevant genes in microarray datasets, and then the selected genes are fed to the wrapper process. Using wrapper (e.g. metaheuristic

algorithms) directly is inefficient due to the high dimensionality of microarray data. Several improved mRMR techniques have been implemented to extend mRMR performance. These improved techniques should be explored further in various type of microarray datasets to find the most informative genes in classifying DNA microarrays.

Acknowledgements
The authors would like to thank the editors and reviewers for their valuable comments. We also would like to thank Universiti Teknologi Malaysia (UTM) for providing Transdisciplinary Research Grant (TDR), Grant No: Q.J130000.3551.05G41

References

[1] Miller S *et al* 2015 *Methods in Microbiology.* **42** 395.
[2] Han J *et al* 2012 *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
[3] Armina R *et al* 2017 *J. of Physics*. 892.
[4] Ding C *et al* 2003 *Comp Sys Bioinf. Proc of the IEEE Bioinfo Conf.* Stanford, CA, USA, **3** 523.
[5] Zhou K Q *et al* 2016 *Art Intel Rev*. **45** 405.
[6] Shayfull Z *et al* 2014 *Advances in Polymer Technology.* **33** 1.
[7] Alshamlan H *et al* 2015 *BioMed Res Int*.
[8] Mohamed N S *et al* 2017 *Exp Sys with App*. **90** 224.
[9] Alomari *O A et al* 2018 *Applied Intelligence*, June.
[10] Angulo A P 2018 *Inf (Switzerland)*, **9** 1.
[11] Wang A *et al* 2015 *J. of Comp Inf Sys.* **5** 1563.
[12] Bolón-Canedo V *et al* 2014 *Inf Sc*. **282** 111.
[13] Zhang Y *et al* 2018 *BioMed Res Int.* 7538204 .
[14] B. Chandra *et al* 2018, *IEEE Conf on Comp Intel in Bioinfo and Comp Bio*. 1–6.
[15] Alshamlan H M *et al* 2015 *Comp Bio and Chem*. **56** 49.
[16] Ramírez-Gallego S *et al* 2017 *Int J. of Intel Sys*. **32** 134.
[17] Kusairi R M *et al* 2017 *Int J. on Adv Sci, Eng and Info Tech*. **7** 1595.
[18] Du W *et al* 2013 *Int J. of Data Mining and Bioinfo*. **7** 58.
[19] Lazim D *et al* 2017 *Art Intel Rev.* 10.1007/s10462-017-9580-4.
[20] A. Deris *et al* 2017 *J. of Physics: Conf Series.* **892** 1.