UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

An Improved Reweighted Recursive Feature Elimination (iRRFE) Approach for Identifying Informative Cancer Genes in Multiclass Classification

Nur Izzati Binti Abd Rahim[1] and Mohd Saberi Bin Mohamad[2]

[1] Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia
[2] Artificial Intelligence and Bioinformatics Group, Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia

izzati.tie03@gmail.com, saberi@utm.my

Abstract. Cancer classification based on microarray gene expression data has advanced substantially in the past decade. Improving the classification accuracy has always been a key factor. Besides, the goal of the research is based on assumption of RRFE is that a gene which is not differentially expressed but connected to other genes. However, the gene expression data often lack of stability. Besides, the genes are assuming to be independent depends on gene expression data. Apart from this issue, an improved Reweighted Recursive Feature Elimination is proposed to overcome the problems in this research. The main result that should produce is on the increasing of the classification accuracy. Other than that, the genes are success to express as well as to compromise with the informative cancer genes.

Keywords: reweighted recursive feature elimination, support vector machine, multiple support vector machine recursive feature elimination.

1 Introduction

The high dimensional data generated through the microarray technology will contain noise or irrelevant genes (Anaissi et al., 2013). These irrelevant genes will affect the classifier in producing classification models with lower accuracy or higher error rates. While, in terms of cancer classification, researchers require precise predictive tools and group of relevant genes for biomarkers (Duval and Hao, 2010). Informative cancer genes are identified based on high throughput gene expression profiling by prognostic gene signatures.

The problem statement in this research is based on the assumption of RRFE that the gene is not differentially expressed but connected to other genes and it expressed that genes should have a higher effect on the classifier compared to those genes that do not have any connections to other deregulated genes. Furthermore, genes which depends on gene expression data, mostly lack of performance, the small number of genes which apply the same method to produce different results of datasets in the gene lists which mostly genes identified by standard method (Ein-Dor et al., 2005).

The objectives is to develop an improved reweighted recursive feature elimination (iRRFE) based on problems which gene is not differentially expressed and the informative cancer which not fully revealed by using supplementary method of multiple support vector machine recursive feature elimination (MSVM-RFE) in order to improve the classification accuracy. Furthermore, the validatation and evaluation for the successful proposed method of an improved reweighted recursive feature elimination (iRRFE) for identifying informative cancer genes in multiclass classification is done. In addition, the scope of this development is the research is done by two main ideas which is explaining the functions of genes and pathways and another else is identification of prognostic and predictive gene signatures. Besides, the reweighted recursive feature elimination (RRFE), which the classification process by changing the notch features criteria of RFE and included the pathway knowledge described by (Guyon et al., 2002). In last, the method come together in this research are Support Vector Machine (SVM) and Google Page Rank to improve the capability and effectiveness throughout this research.

2 An Improved Reweighted Recursive Feature Elimination

2.1 Datasets of Breast Cancer

The datasets used in this research applied to three single datasets as well as an integrated dataset consisting of 588 breast cancer samples; lymph node, gene prognostic signature and estrogen receptor positive breast carcinomas. These cancer datasets are mainly multiclass cancer datasets, which are gene expression datasets obtained through microarray technology.

2.2 Measurement involved

Table 2.1: Differences between measurement

Measurement	Selected Gene	Lowest Error Rates	High Performance	Accuracy Estimation
Leave-one-out error (LOO)	x	x	x	✓
Classification Accuracy	✓	✓	✓	✓
Biological Validation	✓	x	x	x
Number of Genes	x	✓	x	x
Computational Time	x	x	✓	x

Based on table 2.1, it can conclude that, classification accuracy is the best performance measurement to be used. This is because, it considered all the characteristics which are selected genes, lowest error rates, high performance and accuracy estimation.

2.3 Classification Accuracy Formula

The classification accuracy, Acc in this research is calculated based on the following formula:

$$Acc = 1 - err^{B632+}$$

Where err^{B632+2} is the .632+ bootstrap error rates obtained. The range of the classification accuracy is between the value 0 and 1, where the higher or closer the value 1 means the accuracy is better.

2.4 Parameter Setting

The parameter involved in the improved Reweighted Recursive Feature Elimination (iRRFE) has been tabulated according to the parameter name, parameter's value and the description of each parameter. The complete list of parameter involved in iRRFE method has been listed in Table 4.1.The value for each parameter listed in the table is according to the values set by (Johannes, M., et al., 2015), which is known as the default values for varSelRRFE.

2.5 Previous Research and Proposed Research

The implementation of the existing work by (Johannes, M., et al., 2012) which using reweighted recursive feature elimination (RRFE) with a Support Vector Machine Recursive Feature Elimination (SVM-RFE) that produce the classification accuracy as a result. While the Improved Reweighted Recursive Feature Elimination (iRRFE) of the proposed method which using the additional functionalities of Multiple Support Vector Machine Recursive Feature Elimination (MSVM-RFE) to produce and improved the classification accuracy.

3 Result and Discussion

Classification accuracy is the best performance measurement to be used. This is because, it consider all the characteristics which are the good measurement for selected genes, the lowest error rates obtained, high performance of AUC value and a predictable accuracy estimation. The range of the classification accuracy is between the value of 0 and 1, where the higher or closer the value to 1 means the accuracy is better

3.1 Multiple Support Vector Machine Recursive Feature Elimination (MSVM-RFE)

In this improvement for Reweighted Recursive Feature Elimination (RRFE), gene selection and classification is implemented by MSVM-RFE. In order to identifying an informative cancer gene in multiclass classification, the best method can used is Multiple Support Vector Machine Recursive Feature Elimination (MSVM-RFE). This is because MSVM-RFE is a better gene subset and it able to improve classification accuracy. It uses the *extractFeatures()* function to extracts the features which have been selected by the classifiers during the cross validation along with the number of times they have been choosen.

3.2 10-fold Cross Validation

The classification result used 10-fold cross validation in terms of mean AUC and accuracy using conventional RRFE method (rerun "pathClass" package) based on lymph node (GSE2034), gene prognostic signature (GSE6532) and estrogen receptor positive breast carcinoma (GSE7390) which in "pathClass" package. The result in Table 3.1 shows the algorithm involved together with the best C and test AUC value.

Table 3.1: The classification accuracy using 10-fold cross validation based on breast cancer in "*pathClass*" package.

Algorithm	Best C	Best Model Spanbound	C	Features	Test AUC	Finished fold	Models of repeat
RFE	0.1	0	1	7	0.6666667	2	1
RRFE	1	0	100	9	0.9	10	5
Improved RRFE	100	0	1	6	1	10	5

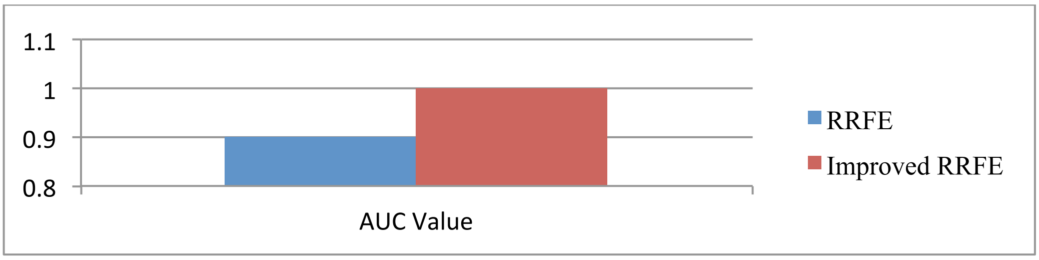

Figure 3.1: Difference AUC value between previous RRFE and Improved RRFE

Table 3.1 shows the classification accuracy using 10-fold cross validation based on breast cancer in "pathClass" package. It complies with the existing RFE method with the previous RRFE and Improved RRFE to compare the best C and test AUC, which method give the best output. As we can observe, as we stressed the AUC value for the main result, we can see that the previous RRFE obtained 0.9 while the Improved RRFE obtained 1. This can be concluded that, the Improved RRFE perform well for AUC value and produced a better classification accuracy compared to RRFE.

While in Figure 3.1, shows the difference AUC value between previous RRFE and Improved RRFE in bar graph for better representation. The previous RRFE indicate 0.9 for value of AUC while the improved RRFE indicate value of 1 for AUC. This represent that improved RRFE produced a better value of AUC compared to previous RRFE.

3.3 Area Under the Curve (AUC)

Based on this improvement of improved reweighted recursive feature elimination (iRRFE), the AUC graphs are represent in box plot graph and ROC curve which involve the true positive and false positive rate.

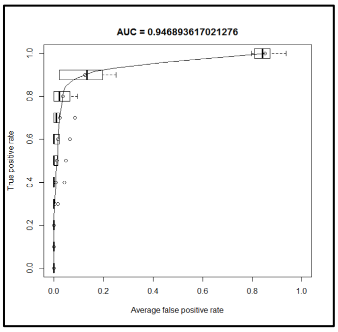

Figure 3.2: The distribution of AUC for each repeat of the cross-validation.

Figure 3.2 shows the box plot of repeated cross validation which produced a better performance after undergoes many times of process the run time. It shows that the value of AUC is near to 1 which can stated that the performance is much better when running in many cycles.

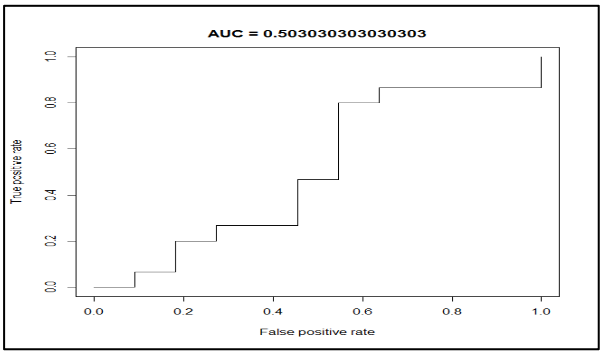

Figure 3.3: ROC curve of AUCs

Figure 3.3 indicate the plot is the ROC curve of the AUCs. The AUC value has been defined to be above than 0.5 to give a better performance. In this graph it shows that the AUC value is greater than the standard. So it proved that the plot RRFE for the improved method is better than the previous RRFE which only plot at 0.393939393939394 that's means under the standard value needed.

3.4 Receiver Operating Characteristic (ROC) Curve

Another performance evaluation is based on Receiver Operating Characteristic (ROC) curve. ROC curve is a plot to evaluate a binary discrimination system. However, the area under curve (AUC) is used as a measurement for performance evaluation.

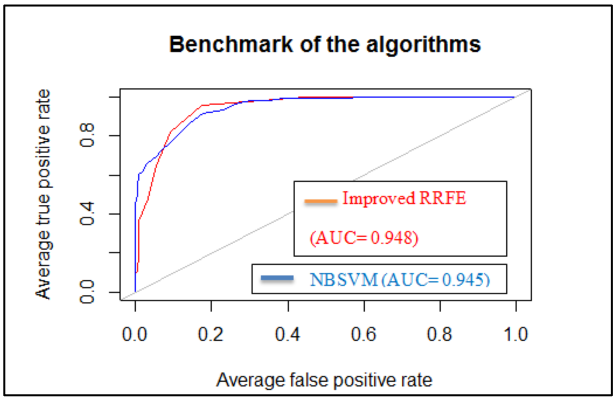

Figure 3.4: ROC curve for all 2 algorithms

Table 3.2: Difference between previous RRFE and Improved RRFE

Algorithm	AUC value
RRFE	0.941
Improved RRFE	0.948

Based on Figure 3.4, it shows the benchmark of the algorithm between RRFE and NBSVM. The AUC value obtained by RRFE is 0.948 while for NBSVM AUC value is 0.945. The value obtained is not too far from each other, but it concluded that, RRFE is more relevant method to be used which can produced a better AUC value which near to 1.

While Table 3.2 indicates the difference between previous RRFE and Improved RRFE which obtained and it shows that the improved RRFE produced a higher AUC value. This is because the improved RRFE gain the AUC value which more closest to 1 compared to previous RRFE.

3.5 Standard Deviation cut off

The standard deviation cut off is used to filtering the genes and observed the remaining genes after undergoes the cut off process. Besides, the inner cross validation will run the process to identify the choosen lambda. The sd.cutoff is a cut off on the standard deviation (sd) of genes. Only genes with sd > sd.cutoff stay in the analysis.

Table 3.3: Differences between sd cutoff of previous RRFE and Improved RRFE

Algorithm	**Sd.cutoff**	**Genes Remained**	**Lambda choosen**
RRFE	150	142	1
Improved RRFE	50	82	0.1

The table 3.3 shows the sd cutoff of previous RRFE and Improved RRFE which has been reduced to produce a good performance for the iteration of the cycle. Besides, the low sd cutoff will indicate the less genes which remained in the analysis. The value for lambda choosen produced a least number of lambda value.

References

1. Anaissi, A., Kennedy, P. J., Goyal, M., and Catchpoole, D. R. (2013). A balanced iterative random random forest for gene selection from microarray data. BMC Bioinformatics, 14(1), 1-10.
2. Duval, B., and Hao, J.-K.(2010). Advances in metaheuristics for gene selection and classification of microarray data. Briefings in Bioinformatics, 11(1), 127-141.
3. Guyon, I. et al. (2002) Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, 46(1), 389-422.
4. Johannes,M. et al. (2011) pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery. Bioinformatics.
5. Johannes, M. et al (2012). pathClass: SVM-based classification with prior knowledge on feature connectivity. German Cancer Research Center Heidelberg, Germany. Version 0.8.0.