Multiclass Prediction for Cancer Microarray Data Using Various Variables Range Reduction Based On Random Forest

Kohbalan Moorthy[1,*], Mohd Saberi Mohamad[1], Safaai Deris[1]

[1]Artificial Intelligence & Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.

kohbalan@gmail.com, saberi@utm.my, safaai@utm.my

Abstract. Continuous data mining has led to the generation of multi class datasets through microarray technology. New improved algorithms are then required to process and interpret these data. Cancer prediction tailored with variable reduction process has shown to improve the overall prediction accuracy. Through variable reduction process, the amount of informative genes gathered are much lesser than the initial data, yet the selective subset present in other methods cannot be fine-tuned to suit the necessity for particular number of variables. Hence, an improved technique of various variable range reduction based on Random Forest method is proposed to allow selective variable subsets for cancer prediction. Our results indicate improvement in the overall prediction accuracy of cancer data based on the improved various variable range reduction technique which allows selective variable reduction to create best subset of genes. Moreover, this technique can assist in variable interaction analysis, gene network analysis, gene-ranking analysis and many other related fields.

Keywords: Variable Reduction, Cancer Prediction, Random Forest, Gene Expression, Microarray Data

1 Introduction

Microarray technology allows continuous analysis and interpretation of the expression levels present in the observed variables from microarray data. Analysing microarray data is a challenging task, as the high dimensionality of the data requires large processing power with sufficient amount of memory resources. Furthermore, microarray technology allows the expansion of information of the sample itself, where detailed insights of the data can be used for gene regulation and identification based on gene expression data [1]. In addition, it has been used in studies related to cancer prediction, identification of relevant variables for diagnosis or therapy and investigation of drug effects on cancer prognosis [2].

Biologists require accurate predictive tools as well as group of relevant variables for biomarkers in cancer identification [3]. Cancer informatics has been expected to be a part of the advancement in the identification and validation of biomarkers through the combine interdisciplinary fields, which expands from the bioinformatics [4]. Prior to cancer prediction, performing variable reduction allows grouping of relevant variables into a subset. Some of the main reasons for performing variable reduction are to avoid over fitting for improved model performance, to gain faster and less costly models and lastly to dig deeper into the data generation processes .

Variable reduction approach is divided into three main categories, which are filter based approach, wrapper based approach and embedded based approach [5]. Filter based approach is defined as when the variable reduction process is carried out independently of the cancer prediction algorithms. If the classifier is being used to evaluate every selected subset of the variable reduction process throughout the entire prediction process, then it is known as a wrapper based approach [6]. Embedded approach uses the same classifier dependent reduction as the wrapper based approach, except that it has better computational complexity. According to Wong, Leckie and Kowalczyk [7], filter based approach performs variable reduction without any dependence on the classifier being chosen, which may not be sufficient enough to generate higher accuracy in cancer prediction as those of wrapper and embedded approaches, which have certain degree of dependencies with the classifier algorithm being used. In spite of that, wrapper based approach is not preferred in sample prediction due to huge combination of variable subset required to be examined. Moreover, the wrapper method requires high computation time and it is much slower in determining the best subset of variables [8].

Accurately categorizing the selected variables into their respective class as into normal or tumour is known as the process of binary prediction. Classifier can be defined as an artificial intelligence device, which has the potential to make prediction. In usual cancer prediction scenario, most developed algorithms focus on maximizing the overall correct classifiers in order to gain higher prediction accuracy even though there is an imbalance in the different class size [9]. Some examples of classifiers are support vector machines (SVM), neural network (NN), k-nearest neighbor (kNN) and classification tree.

In genetic associated studies, Random Forest has been used widely for both prediction and variable reduction [10]. Random Forest was first developed by Breiman [11] for the purpose of classification, regression, clustering and also survival analysis. In this field, the practice and application of variable ranking are according to the variables contribution towards a disease. Random forest has been one of the favoured methods used in variable importance measurement for variable ranking and selection. Diaz-Uriarte and Alvarez de Andres [12] had proposed a variable selection and classification based on Random Forest for the first time as an embedded approach. Besides that, Random Forest algorithm is effective in predicting samples, as well as revealing interactions among the variables. Additionally, a limiting value is achieved as the number of trees set in the Random Forest is increased continuously, making it an ideal error classifier with no over fitting occurrence of the data. In Random Forest, trees are

grown, and from the training sample, each tree grows without pruning from the actual data based on random variable reduction.

For the creation of gene expression profiles, many researchers are continuously seeking for state of the art prediction algorithms that can provide better accuracy. Variable reduction has played a vital role in increasing the prediction accuracy for cancer related disease but most of the variable reduction techniques available are unrelated to the prediction algorithm. Moreover, the amount of variables selected in variable sub-sets are dependent on the variable reduction technique used and cannot be fine-tuned to suit the requirement for particular number of variables. Hence, we propose a technique known as various variables range reduction based on a Random Forest method for selective subset, leading to better prediction of cancer datasets.

In this article, we begin by describing the methodology section where the proposed technique is briefly explained; followed by the result and discussion section, where the main characteristics of the datasets are explained, and the complete analysis of the findings is presented. Comparisons with previous similar research papers are also presented to further justify the improvement achieved using the proposed technique. Lastly, the future works and conclusion of this article are presented.

2 Methodology

Diaz-Uriarte and Alvarez de Andres [12] first proposed the variable selection through Random Forest algorithm. Moorthy and Mohamad [13] then proposed an improved version of the variable selection. In this research, we propose an improvement on the variable reduction technique based on the Random Forest method, which are various variables range reduction. Most existing techniques and methods used for variable reduction do not reveal the amount of variables selected for training the classifier. Moreover, the selected subset of variables is very dependent on the variable reduction technique and does not have the capability to tune and finalize the amount of the selected variables for extended usage in other related fields, such as gene network analysis, gene-gene interaction analysis, and gene annotations. Besides that, most of the variable reduction techniques produce constant output of variables for the use of the prediction algorithms. Therefore, there are no possibilities of tweaking that particular variable reduction technique to evaluate the different output performance of the classifier.

Through this research, an enhancement to the variable reduction technique is introduced to provide the flexibility and options to generate different variable sets with better accuracy, as well as the ability to control the amount of variables required on each variable subset. The idea of this improvement focuses on allowing the variable reduction algorithm to test and evaluate a certain range of variables from the overall dataset and evaluate the final prediction accuracy. Furthermore, it allows analysis and comparison of different variable subsets towards the prediction accuracy. The main reason for introducing this improved variable reduction technique is to provide various variables range reduction in any particular selected variable subset for better prediction of cancer. Moreover, it is also to allow other researchers to further tweak and

select their desire range of variables in any particular variable subset, which can provide better analysis capability in other research areas.

In order to achieve the proposed various variable range technique, modification to the steps in the backward elimination process were carried out to accept inputs of selective range of variables, which were taken as minimum value (MinVar) and maximum value (MaxVar). Prior to that, the cancer dataset were represented in two different forms of dataset information (Data) and to class the dataset to (Class). While performing the backward elimination process, a new subset was generated and evaluated where the previous error rates obtained (p.mean) were compared with the current error rates obtained (c.mean), and if there was a lower error rates, then the previous best would be replaced with the current best. Once the best subset of variables was determined and the required number of variables was satisfied, we then used the variable subset (bestSub) for the prediction process. A complete flow of the various variable range reduction technique is been presented in Figure 1, where the dotted line represents the changes made to achieve the range selection.

```
Various Variables Range Reduction Technique
1:  Input: Data, Class, MinVar and MaxVar
2:  Output: Selected variables and error rates
3:  while backward elimination process = true do
4:     removes fraction of variables;
5:     test and evaluate remaining variables;
6:     c.mean = current error rates;
7:     p.mean = previous error rates;
8:     if c.mean <= p.mean
9:        p.mean = c.mean;
10:       selVar = current subset of variables;
11:       if selVar <= MaxVar and selVar >= MinVar
12:          bestSub = selVar;
13:       end if
14:    end if
15:    if selVar < MinVar
16:       break;
17:    end if
18: end while
```

Fig. 1. Pseudo code for the various variables range reduction technique developed for controlled amount of selected variables in a particular subset.

3 Results and Discussion

In this research, we used cancer related datasets, which were gene expression dataset obtained through the microarray technology. The datasets involved in this research could be grouped into various cancer types, which include breast cancer (Breast), blood cancer (Lymphoma), small round blue cell tumours (SRBCT), brain cancer

(Brain) and a set of 60 human tumour cell lines derived from various tissues of origin (NCI60). These cancer datasets were multiclass cancer datasets where various cancer types are considered simultaneously in the microarray experiments.

The cancer datasets used for this research were in text file format and had been pre-formatted to suit the software. For each of the cancer dataset, they have two main text files, which were class file and data file. The class file contained the information to identify the data file according to normal or tumour samples. The data file consists of numerical values, where the rows represent the total number of variables in any particular cancer dataset and the columns represent the total number of patients. The detailed description of the cancer dataset is presented in the Table 1, where the number of variables, patients and the main reference of the data are listed.

Table 1. Main characteristics of the cancer dataset used in this research.

Dataset Name	Variables	Patients	Class	Reference
Breast	4869	95	3	[14]
Lymphoma	4026	62	3	[15]
SRBCT	2308	63	4	[16]
Brain	5597	42	5	[17]
NCI60	5244	61	8	[18]

The complete analysis for the selected cancer datasets had been tabulated according to selected various variables range reduction, and both the number of variables in a subset and error rates were obtained. The selected various variables range had been set to into four different partitions as to 2 to 10 variables for the first range, 10 to 50 variables for the second range, 50 to 250 variables for the third range and the final range from 250 variables to the maximum number of variables present in any particular dataset.

The selected various variables range reduction settings executed were used to determine the local optimum variables subset for the entire dataset and each subset could be selected to be further used into the prediction process. In terms of the error rates calculation, the .632+ Bootstrap error rates from Efron and Tibshirani [19] had been applied. The complete result is presented in Table 2.

From the results gathered, we can see that the best subset of variables for Breast dataset consists of 214 variables that make up the lowest error rates obtained compared to other various variables range reduction, but the recommended subset would be 6 variables which resulted in an error rates of 0.349810. This is because the difference in the total selected variables differs from 6 variables to 214 variables which is an increase of 36 folds higher whereas the differences in the error rates were merely 2%. This 2% differences could not compensate to the improvement in accuracy compared to the ratio of the variables.

Apart from that, the Lymphoma and NCI60 dataset showed a similar variables range category as both the datasets has a best subset between 50 to 250 variables range.

Table 2. Prediction error rates of the cancer dataset based on various variables range reduction technique where the shaded area represents lowest error rates.

Various Variable Range Reduction	Breast		Lymphoma		SRBCT		Brain		NCI60	
	*No of Variables	Error Rates	*No of Variables	Error Rates	*No of Variables	Error Rates	*No of Variables	Error Rates	*No of Variables	Error Rates
2 – 10	6	0.349810	2	0.039340	9	0.041312	9	0.205099	10	0.331037
10 – 50	45	0.357240	30	0.041338	22	0.044975	18	0.227755	30	0.338226
50 – 250	70	0.347039	73	0.038521	52	0.041826	42	0.211870	60	0.320460
250 – max**	214	0.342566	222	0.043340	248	0.044492	246	0.218365	230	0.335289

* Total variables present in any particular selected subset.
** All variables in the dataset.

The best subset for Lymphoma dataset consists of 73 variables where the error rates obtained is 0.038521 whereas for NCI60 dataset, the best subset is 60 variables with error rates of 0.320460. These datasets requires larger number of variables to achieve higher prediction accuracy since the dataset is a multiclass dataset and certain minimum informative variables is required to identify and train the classifier to predict the different number of class present in those datasets, especially for NCI60 dataset where there are 8 different types of tumours present in that dataset.

SRBCT and Brain cancer dataset requires lesser variables in the subset to achieve lowest error rates in its various variables range reduction category compared to other multiclass dataset. These both datasets has the best subset of 9 variables which is the most minimum number of variables required to achieve the high accuracy in the prediction. The error rates obtained are 0.041312 and 0.205099 for SRBCT and Brain cancer dataset respectively. Most probably, both this datasets had much lesser informative variables in overall compared to other datasets. Therefore, higher number of variables would only affect the prediction accuracy and increased the error rates. With the various variables range reduction, the best subset from each range partition had been used for the random forest classifier to obtain the highest possible accuracy, which is presented in the Figure 2.

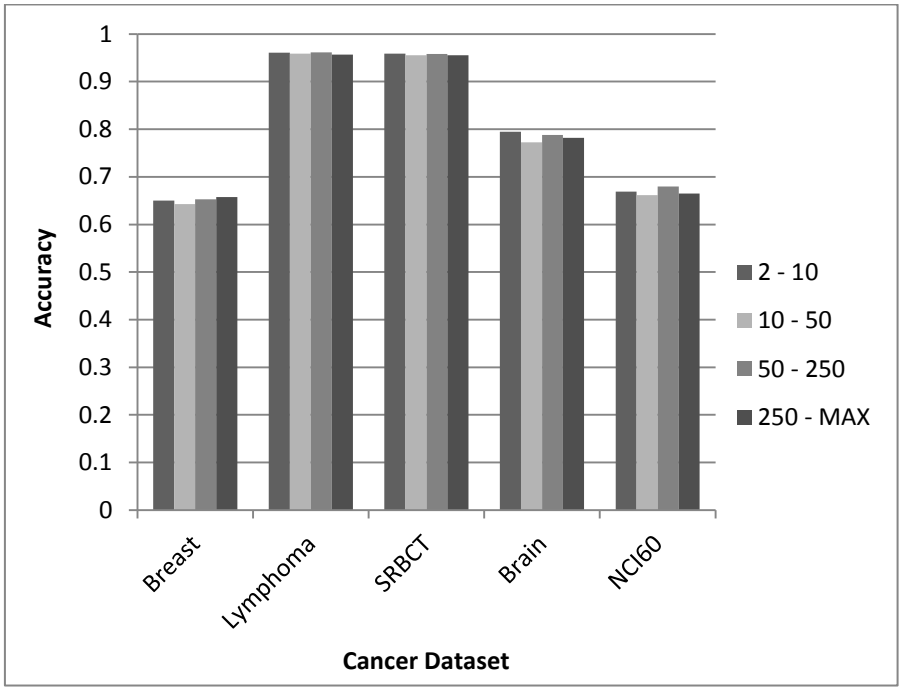

Fig. 2. Comparison of different various variables range reduction towards the overall prediction accuracy of the cancer datasets.

From our analysis, we could deduce that the suitable range for informative variables was at 5 – 75 variables, as most of the dataset shown better or higher accuracy in this range. Even though the difference was not intermittent in terms of accuracy, but the amount of variables were either too less or too many for other selected ranges. However, other researchers may use the variance of the variables amount for subsequent analysis as well as a variable filtration for large datasets. Besides that, the various variables range reduction can be altered to suit other requirements such as for the construction of gene network analysis, genes functional annotation through gene ontology and many more subsequent analyses.

In order to justify the improvement achieved using this improved variable reduction technique, a comparison with a previous work was done and the accuracy achieved is shown in Figure 3. Based on the comparison, we can see that our improved technique has increased the prediction accuracy for all the datasets used. The average improvement between our results and previous work prediction accuracy for all datasets is 5.29%, where there is a 3.69% increase in prediction accuracy for Breast cancer dataset, a huge 7.64% increase for Lymphoma dataset, 3.39% increase for SRBCT dataset, 3.97% increase for Brain dataset and finally a 7.75% increase in prediction for NCI60 dataset. This is due to the fact that the selected variables in the variable reduction process have more significant variables compared to the previous work.

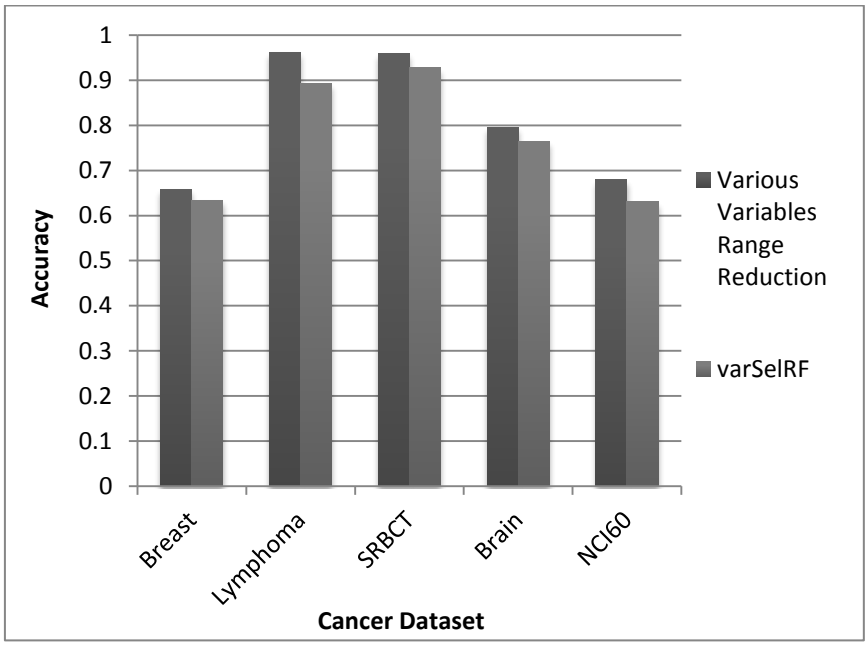

Fig. 3. Prediction accuracy comparison of our improved technique (Various Variables Range Reduction) with previous works (varSelRF) from Diaz-Uriarte and Alvarez de Andres [12].

4 Future Works

Cancer detection through Single nucleotide polymorphism (SNP) is a crucial stage in the prediction of cancer patients and it would be another step of advancement if the Random Forest method can be altered to accept feeds from the SNP type microarray data in future. Besides that, the annotation of the selected variables and cross-referencing with genes databases could provide better understanding and validation of future predicted variables subsets.

5 Conclusion

The various variables range reduction technique has been tested with five different multiclass cancer datasets and the outcome of the prediction has been presented in the result and discussion section. With the wide possibilities of variables subset selection, the accuracy of the prediction based on the selected subsets has shown similar or better accuracy with no such fluctuation on the overall accuracy. This allows different range of variables to be selected from the entire datasets without deteriorating the prediction accuracy.

Most variable reduction techniques do not provide the actual number of variables in the selected subset, nor the flexibility to tune the amount of variables to be chosen in any particular variable subset prior to prediction. We have shown a method of solution with the proposed various variables range reduction technique, which allows fine-tuning of the amount of variables selected in any particular variable subset without degrading the prediction accuracy. Through the development of the various variables range technique for the Random Forest variable reduction, different subsets of variables with better prediction accuracy have been listed for various use of gene expression analysis. The possibility for further analysis through gene network analysis, gene – gene interaction analysis and other related analysis is also made available, for the researchers may have their own preference of range of selection to obtain various sets of variables. This will not only allow controlling the amount of variables to be obtained but also provide accuracy of estimation based on the comparison of the selected variables.

Acknowledgements. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Paz, J.L., Seeberger, P.H.: Recent Advances and Future Challenges in Glycan Microarray Technology Carbohydrate Microarrays. In: Chevolot, Y. (ed.), vol. 808, pp. 1-12. Humana Press (2012)
2. Liew, A.W.-C., Law, N.-F., Yan, H.: Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 12, 498-513 (2011)
3. Duval, B., Hao, J.-K.: Advances in metaheuristics for gene selection and classification of microarray data. Briefings in Bioinformatics 11, 127-141 (2010)
4. Wu, D., Rice, C., Wang, X.: Cancer bioinformatics: A new approach to systems clinical medicine. BMC Bioinformatics 13, 71 (2012)
5. Van Steen, K.: Travelling the world of gene–gene interactions. Briefings in Bioinformatics 13, 1-19 (2012)
6. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. 42, 409-424 (2009)
7. Wong, G., Leckie, C., Kowalczyk, A.: FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number. Bioinformatics 28, 151-159 (2012)
8. Nanni, L., Brahnam, S., Lumini, A.: Combining multiple approaches for gene microarray classification. Bioinformatics 28, 1151-1157 (2012)
9. Lin, W.-J., Chen, J.J.: Class-imbalanced classifiers for high-dimensional data. Briefings in Bioinformatics (2012)
10. Boulesteix, A.-L., Bender, A., Lorenzo Bermejo, J., Strobl, C.: Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Briefings in Bioinformatics 13, 292-304 (2012)
11. Breiman, L.: Random Forests. Mach. Learn. 45, 5-32 (2001)
12. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
13. Moorthy, K., Mohamad, M.S.: Random forest for gene selection and microarray data classification. Bioinformation 7, 142-146 (2011)
14. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530-536 (2002)
15. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown,

P.O., Staudt, L.M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511 (2000)
16. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7, 673-679 (2001)
17. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436-442 (2002)
18. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O.: Systematic variation in gene expression patterns in human cancer cell lines. Nature genetics 24, 227-235 (2000)
19. Efron, B., Tibshirani, R.: Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 92, 548-560 (1997)