Multiple Gene Sets for Cancer Classification Using Gene Range Selection Based on Random Forest

Kohbalan Moorthy*, Mohd Saberi Bin Mohamad, and Safaai Deris

Artificial Intelligence & Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
kohbalan@gmail.com, {saberi,safaai}@utm.my

Abstract. The advancement of microarray technology allows obtaining genetic information from cancer patients, as computational data and cancer classification through computation software, has become possible. Through gene selection, we can identify certain numbers of informative genes that can be grouped into a smaller sets or subset of genes; which are informative genes taken from the initial data for the purpose of classification. In most available methods, the amount of genes selected in gene subsets are dependent on the gene selection technique used and cannot be fine-tuned to suit the requirement for particular number of genes. Hence, a proposed technique known as gene range selection based on a random forest method allows selective subset for better classification of cancer datasets. Our results indicate that various gene sets assist in increasing the overall classification accuracy of the cancer related datasets, as the amount of genes can be further scrutinized to create the best subset of genes. Moreover, it can assist the gene-filtering technique for further analysis of the microarray data in gene network analysis, gene-gene interaction analysis and many other related fields.

Keywords: Gene Selection, Cancer Classification, Random Forest, Gene Expression, Microarray Data.

1 Introduction

Vast amount of data generation have led to the development of many sophisticated methods and tools for visualization and analysis of data. These huge amounts of data, particularly for the biological analysis and interpretation, are made available through microarray technology [1]. Microarray technology allows continuous analysis and interpretation of the expression levels present in the observed genes from microarray data. Analyzing 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 [2]. In addition, it has been used in studies

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 385–393, 2013.

related to cancer classification, identification of relevant genes for diagnosis or therapy and investigation of drug effects on cancer prognosis [3].

Biologists require accurate predictive tools as well as group of relevant genes for biomarkers in cancer identification [4]. 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 [5]. Prior to classification, performing gene selection allows grouping of relevant genes into a subset. Some of the main reasons for performing gene selection 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 [6].

Gene selection approach is divided into three main categories, which are filter based approach, wrapper based approach and embedded based approach [7]. Filter based approach is defined as when the gene selection process is carried out independently of the classification algorithms. If the classifier is being used to evaluate every selected subset of the gene selection process throughout the entire classification process, then it is known as a wrapper based approach [8]. Embedded approach uses the same classifier dependent selection as the wrapper based approach, except that it has better computational complexity. According to Wong, Leckie and Kowalczyk [9], filter based approach performs gene selection without any dependence on the classifier being chosen, which may not be sufficient enough to generate higher accuracy in classification 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 classification due to huge combination of genes subset required to be examined. Moreover, the wrapper method requires high computation time and it is much slower in determining the best subset of genes [10].

Accurately categorizing the selected genes into their respective class as into normal or tumor is known as the process of binary classification. Classifier can be defined as an artificial intelligence device, which has the potential to make classification [11]. In usual classification scenario, most developed algorithms focus on maximizing the overall correct predictors in order to gain higher accuracy in classification even though there is an imbalance in the different class size [12]. 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 classification and gene selection [13]. Random Forest was first developed by Breiman [14] for the purpose of classification, regression, clustering and also survival analysis. In this field, the practice and application of gene ranking are according to the genes contribution towards a disease. Random forest has been one of the favored methods used in gene importance measurement for gene ranking and selection. Diaz-Uriarte and Alvarez de Andres [15] had proposed a gene 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 genes. Additionally, a limiting value is achieved as the number of trees set in the Random Forest is increased continuously, making it an ideal error predictor 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 gene selection.

For the creation of gene expression profiles, many researchers are continuously seeking for state of the art classification algorithms that can provide better accuracy. Gene selection has played a vital role in increasing the classification accuracy for cancer related disease but most of the gene selection techniques available are unrelated to the classification algorithm. Moreover, the amount of genes selected in gene sub-sets are dependent on the gene selection technique used and cannot be fine-tuned to suit the requirement for particular number of genes. Hence, we propose a technique on gene range selection based on a Random Forest method for selective subset, leading to better classification 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 [15] first proposed the gene selection through Random Forest algorithm. Moorthy and Mohamad [16] then proposed an improved version of the gene selection. In this research, we propose an improvement on the existing gene selection technique based on the Random Forest method, which is gene range selection. Most existing techniques and methods used for gene selection do not reveal the amount of genes selected for training the classifier. Moreover, the selected subset of genes is very dependent on the gene selection technique and does not have the capability to tune and finalize the amount of the selected genes for extended usage in other related fields, such as gene network analysis, gene-gene interaction analysis, and gene annotations. Besides that, most of the gene selection techniques produce constant output of genes for the use of the classification algorithms. Therefore, there are no possibilities of tweaking that particular gene selection technique to evaluate the different output performance of the classifier.

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

In order to achieve the proposed gene section technique, modification to the steps in the backward elimination process were carried out to accept inputs of selective range of genes, 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 back-ward 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 reduction, then the previous best would be replaced with the current best. Once the best subset of genes was determined and the required number of genes was satisfied, we then used the gene subset (bestSub) for the classification process. A complete flow of the gene range technique is been presented in Figure 1, where the dotted line represents the changes made to achieve the range selection.

```
Gene Range Technique
1:  Input: Data, Class, MinVar and MaxVar
2:  Output: Selected genes and error rates
3:  while backward elimination process = true do
4:     removes fraction of genes;
5:     test and evaluate remaining genes;
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 genes;
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 gene range selection developed for controlled amount of selected genes 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 includes adenocarcinoma, breast cancer, colorectal cancer, leukemia and prostate cancer. These cancer datasets were primarily binary, which are known as two-class dataset and consists of both the normal, and tumor based patient samples.

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 tumor samples. The data file consists

of numerical values, where the rows represent the total number of genes 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 genes, patients and the main reference of the data are listed.

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

Dataset Name	Genes	Patients	Reference
Adenocarcinoma	9868	76	[17]
Breast	4869	77	[18]
Colon	2000	62	[19]
Leukemia	3051	38	[20]
Prostate	6033	102	[21]

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

The minimum of two genes were preset, as each gene from the tumor and normal was required as the minimum informative genes for the classification purpose. The selected gene range settings executed were used to determine the local optimum genes subset for the entire dataset and each subset could be selected to be further used into the classification process. In terms of the error rates calculation, the .632+ Bootstrap error rates from Efron and Tibshirani [22] had been applied. The complete result is presented in Table 2.

From the results obtained, we can see that the best number of genes for Adenocarcinoma dataset was 12 genes, with the classification error rates obtained as low as 0.1801 compared to other selected range. Even though with 222 genes, the error rates obtained were the lowest among all the selected range, which was 0.1745, the number of genes in this particular subset was huge and did not compensate the 3% improvement in accuracy compared to the ratio of the genes. This could be the indication that there were some genes in the subset of 222 genes, which might be useful in increasing the overall accuracy. Similar case happened to the Breast cancer dataset as we can see that the lowest error rates obtained were 0.3249 with 214 genes in the selected subset. The recommended subset of genes for this dataset would be 56 genes as the error rate obtained was 0.3257, which was similar but slightly higher than the best error rate and the difference was only 0.2%, yet the difference in the number of genes is 214 genes over 56 genes.

Apart from that, Colon cancer dataset and Leukemia dataset showed a similar gene range selection, as the best gene subset for Colon cancer dataset consisted of 18 genes whereas 9 genes were obtained for the Leukemia dataset. The lowest error rates obtained for Colon cancer dataset and Leukemia dataset were 0.1539 and 0.0753, respectively. Most probably, both this datasets had much lesser informative genes in overall compared to other datasets. Therefore, higher number of genes would only affect the classification accuracy and increased the error rates. For Prostate cancer

Table 2. Classification error rates of the cancer dataset based on gene range selection technique where the shaded area represents lowest error rates

Gene Range	Adenocarcinoma		Breast		Colon		Leukemia		Prostate	
	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates	*No of Genes	Error Rates
2 – 10	2	0.1871445	6	0.3287878	3	0.1655999	2	0.07934855	2	0.06249576
10 – 50	12	0.1801157	29	0.3311092	18	0.1539477	9	0.07529078	18	0.06060309
50 – 250	58	0.1844413	56	0.3256622	56	0.1681416	44	0.08684319	109	0.06014227
250 – max**	222	0.1744816	214	0.3249034	214	0.1625876	210	0.08636869	212	0.05849378

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

dataset, the average error rate obtained was 0.06 and with the different gene range selection, there were no significant differences. Even though the best gene subset contained 212 genes, based on the error rates differences, the preferred gene subset would be with 18 genes. This could be due to the amount of neutral genes, which did not contribute enough in the classification. With the various selection ranges, 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.

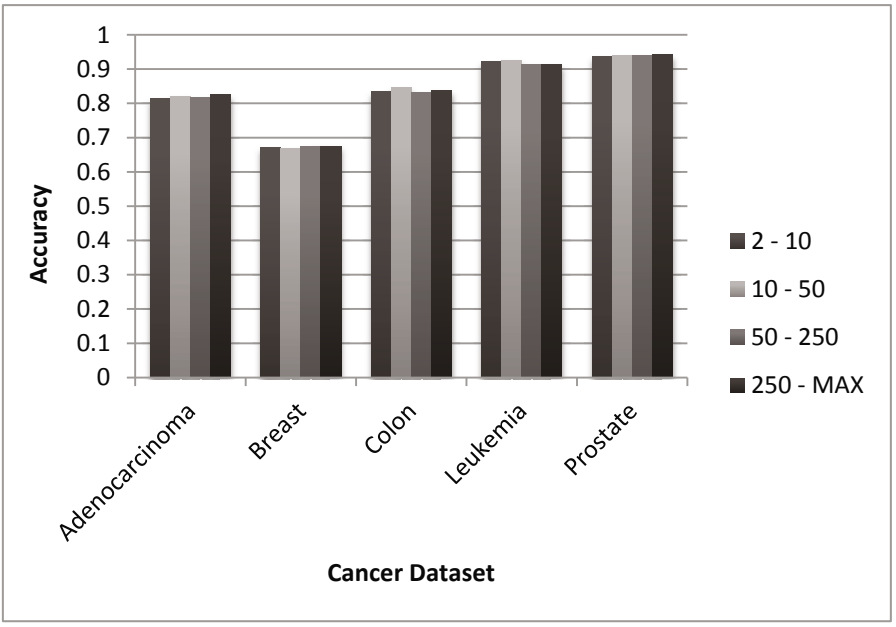

Fig. 2. Comparison of different gene range selection towards the overall classification accuracy of the cancer datasets

From our analysis, we could deduce that the suitable range for informative genes was at 10 – 50 genes range, 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 genes were either too less or too many for other selected ranges. However, other researchers may use the variance of the genes amount for subsequent analysis as well as a gene filtration for large datasets.

Besides that, the gene range selection 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.

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 genes and cross-referencing with genes databases could provide better understanding and validation of future predicted gene subsets.

5 Conclusion

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

Most gene selection techniques do not provide the actual number of genes in the selected subset, nor the flexibility to tune the amount of genes to be chosen in any particular gene subset prior to classification. We have shown a method of solution with the proposed gene range selection technique, which allows fine-tuning of the amount of genes selected in any particular gene subset without degrading the classification accuracy. Through the development of the gene range technique for the Random Forest gene selection, different subsets of genes with better classification 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 genes. This will not only allow controlling the amount of genes to be obtained but also provide accuracy of estimation based on the comparison of the selected genes.

Acknowledgements. Institutional Scholarship MyPhd provided by the Ministry of Higher Education of Malaysia finances this work. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot numbers: QJ130000.7123.00H67 and QJ130000.7107.01H29).

References

1. Paz, J.L., Seeberger, P.H.: Recent Advances and Future Challenges in Glycan Microarray Technology. In: Chevolot, Y. (ed.) Carbohydrate Microarrays, vol. 808, pp. 1–12. Humana Press (2012)
2. Pham, T.D., Wells, C., Crane, D.I.: Analysis of Microarray Gene Expression Data. Current Bioinformatics 1, 37–53 (2006)
3. 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)
4. Duval, B., Hao, J.-K.: Advances in metaheuristics for gene selection and classification of microarray data. Briefings in Bioinformatics 11, 127–141 (2010)
5. Wu, D., Rice, C., Wang, X.: Cancer bioinformatics: A new approach to systems clinical medicine. BMC Bioinformatics 13, 71 (2012)

6. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517 (2007)
7. Van Steen, K.: Travelling the world of gene–gene interactions. Briefings in Bioinformatics 13, 1–19 (2012)
8. 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)
9. 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)
10. Nanni, L., Brahnam, S., Lumini, A.: Combining multiple approaches for gene microarray classification. Bioinformatics 28, 1151–1157 (2012)
11. Asyali, M.H., Colak, D., Demirkaya, O., Inan, M.S.: Gene Expression Profile Classification: A Review. Current Bioinformatics 1, 55–73 (2006)
12. Lin, W.-J., Chen, J.J.: Class-imbalanced classifiers for high-dimensional data. Briefings in Bioinformatics (2012)
13. 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)
14. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001)
15. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
16. Moorthy, K., Mohamad, M.S.: Random forest for gene selection and microarray data classification. Bioinformation 7, 142–146 (2011)
17. Ramaswamy, S., Ross, K.N., Lander, E.S., Golub, T.R.: A molecular signature of metastasis in primary solid tumors. Nature Genetics 33, 49–54 (2003)
18. 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)
19. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96, 6745–6750 (1999)
20. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
21. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209 (2002)
22. Efron, B., Tibshirani, R.: Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 92, 548–560 (1997)