International Journal of Engineering & Technology, 7 (2.15) (2018) 68-71

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Filter-Based Gene Selection Method for Tissues Classification on Large Scale Gene Expression Data

Farzana Kabir Ahmad*, Yuhanis Yusof, Nooraini Yusoff

School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
**Email: farzana58@uum.edu.m*

Abstract

DNA microarray technology is a current innovative tool that has offers a new perspective to look sight into cellular systems and measure a large scale of gene expressions at once. Regardless the novel invention of DNA microarray, most of its results relies on the computational intelligence power, which is used to interpret the large number of data. At present, interpreting large scale of gene expression data remain a thought-provoking issue due to their innate nature of "high dimensional low sample size". Microarray data mainly involved thousands of genes, n in a very small size sample, p. In addition, this data are often overwhelmed, over fitting and confused by the complexity of data analysis. Due to the nature of this microarray data, it is also common that a large number of genes may not be informative for classification purposes. For such a reason, many studies have used feature selection methods to select significant genes that present the maximum discriminative power between cancerous and normal tissues. In this study, we aim to investigate and compare the effectiveness of these four popular filter gene selection methods namely Signal-to-Noise ratio (SNR), Fisher Criterion (FC), Information Gain (IG) and t-Test in selecting informative genes that can distinguish cancer and normal tissues. Two common classifiers, Support Vector Machine (SVM) and Decision Tree (C4.5) are used to train the selected genes. These gene selection methods are tested on three large scales of gene expression datasets, namely breast cancer dataset, colon dataset, and lung dataset. This study has discovered that IG and SNR are more suitable to be used with SVM while IG fit for C4.5. In a colon dataset, SVM has achieved a specificity of 86% with SNR while and 80% for IG. In contract, C4.5 has obtained a specificity of 78% for IG on the identical dataset. These results indicate that SVM performed slightly better with IG pre-processed data compare to C4.5 on the same dataset.

Keywords: *Bioinformatics; Feature Selection; High Dimensional Data; Support Vector Machine*.

1. Introduction

Cancer is a complex disease that has been extensively studied over the past decades [1]. However, due to morphological and genetic heterogeneity of the disease, many patients still believe cancer is a deadly illness. Furthermore, it is very difficult to differentiate patients since patients who undergo similar regimen treatments may develop different clinical outcome.

Over a decade, a number of computational methods have been extensively studies to address the heterogeneity issues of cancers in offering better assessment of cancer diagnosis. In the early days, computational techniques such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree and other have been used to predict cancer relapse, which other applied these techniques to determine patients' survivability [2]. Regardless the extensive work in this domain area, the performance of computational models in assessing cancer recurrence is still uncertain and doubtful. Many researchers believe an additional work to probe deeper into cellular mechanisms is required since this devastating disease mainly initiated from accumulation of gene alterations that disrupt the normal cellular activities and ultimately cause tumor growth.

Currently, the innovation of DNA microarray technology has offered massive opportunities for scientists to look deeper into cellular interactions and analyze large scale of gene expression data to determine the state of cancer progression. This new invention of biological technology has not only brought a great impact toward biological domain but has given immense challenges to computer scientists in order to handle big data issues. Hence, even microarray technology offers a new platform in investigating cancer, analyzing large-scale of gene expression data generated by this device is not an easy task [3].

Microarray data analysis is a fast-growing since this device can be used to measure global expression of genes simultaneously. However, analyzing and interpreting large scale of gene expression data remain a challenging issue due to their innate nature of "high dimensional low sample size" [3,4]. Microarray data mainly involved thousands of genes, n in a very small size sample, p. In addition, this data is often overwhelmed, over fitting and confused by the complexity of data analysis [5,6]. Due to the nature of this microarray data, it is common that a large number of genes are may not informative for classification because these genes can be either irrelevant or redundant. For such a reason, feature selection methods also known as gene selection methods become apparently need for both domains: biology and computation/statistics.

This study aims to explore filter gene selection methods in determining the informative genes which are most predictive to its related class for tissue classification. Four popular filter gene selection methods namely Signal-to-Noise ratio (SN), Fisher Criterion (FC), Information Gain (IG) and t-Test will be examined using two common classifiers, SVM and C4.5. These gene selection methods are tested on three large scales of gene expression datasets, namely breast cancer dataset, colon dataset, and lung dataset. In section 2, the four filter-based gene selection algorithms along with SVN and C4.5 classifiers are described. Experimental results and its

corresponding discussion are presented in section 3. Finally, concluding remarks is given in section 4.

2. Filter-based gene selection methods for large scale gene expression data

Tissues classification is an important step to differentiate cancerous and non-cancerous cells in which appropriate treatments can be prescribed to patients. In such attempts, gene selection methods are considered as an essential phase to select optimal features/gene sets can be further used to classify the human tissues to cancerous or non-cancerous group. However, so far only a few computational studies have compared the performance of gene selection method. Hence this study aims to evaluate the performance of gene filtering-based methods to select informative gene that could distinguish class related to tissues classification. In this study, filter-based gene selection is used as it offers easy ways to measure optimal genes, able to rank the selected genes and provide less computational time.

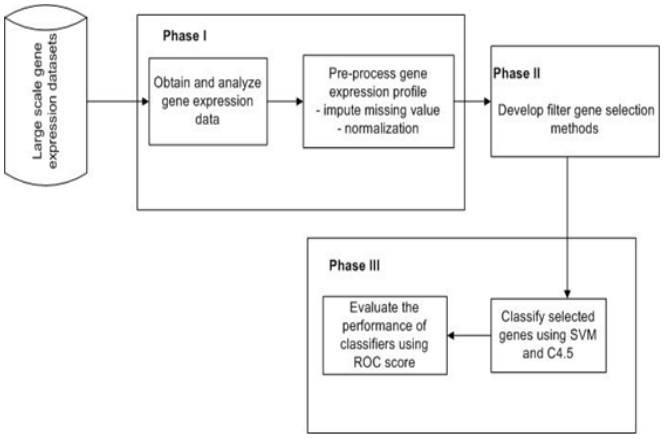

Fig 1: Research framework

The research framework as shown in Figure 1 is constructed to achieve the stated aim. This research framework consists of three main phases, which are; a) Phase I - Data preprocessing phase, b) Phase II - Gene selection phase, and c) Phase III - Validation phase. Each of these phases is explained in the following sub-sections. On the other hand, Figure 2 illustrated the research activities that are initiated to compare the filter-based gene selection methods.

2.1. Data pre-processing phase

Data pre-processing is the initial step in this research, in which obtained datasets undergo the process of imputation to address the missing values issue. The proposed method is performed and tested on three large scales of gene expression datasets, namely breast cancer dataset, colon dataset, and lung dataset. These datasets are obtained from various sources available in National Cancer Institute and the parameter of these datasets is as given in Table 1. 3.

Table 1: Properties of gene expression dataset tested

Dataset	Number of genes	Number of tissue samples
Breast cancer	24,481	32
Colon cancer	6,500	62
Lung cancer	12,533	181

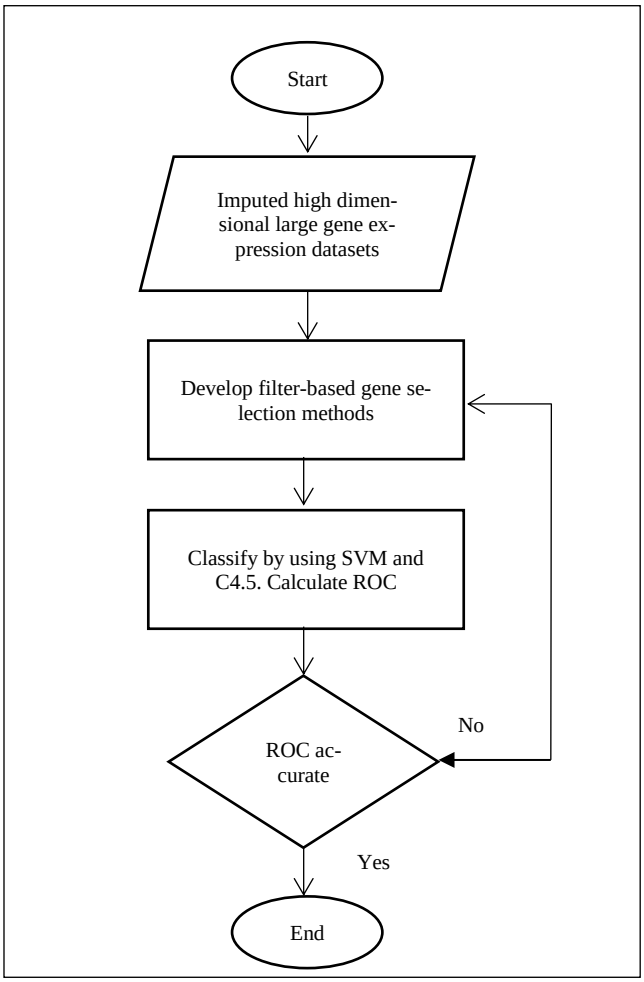

Fig 2: Flow chart of research activities

Prior the subsequent analysis, missing values in gene expression datasets is addressed by using k-nearest neighbor (KNN). This technique has been proposed by many researchers and proven to be robust [7] in imputing missing values [8,9,10]. Furthermore, KNN less affects the performance of downstream analysis. Hence, KNN with Euclidean distance and weighted neighbor genes are used to calculate the missing values in this study

2.2 Filter-based gene selection phase

Microarray measures thousands of gene expression profiles that commonly contain redundant and non-informative genes that may jeopardize the performance of any standard classifiers in the downstream analysis. Hence, gene selection methods have become a necessity to reduce the complex time processing and enhance the accuracy of classifier. In this study, four filter filter-based gene selection techniques namely SNR, FC, IG and t-Test are used to select the informative genes before classifying them in the SVM and C4.5 classifier. The following sections described in detail of these techniques.

2.2.1 Signal-to-noise ratio

SNR is a statistical gene selection method that measures the mean and standard deviation of gene that could be used to distinguish tissues sample into a particular class in respect to another class. The formula of SNR used in this study is as given below:

$$\mu_{class1} - \mu_{class2} / \delta_{class1} - \delta_{class2} \quad (1)$$

where μ is the mean and δ is standard deviation of tissue sample for class 1 = cancerous group and class 2 = non-cancerous group

2.2.2 Fischer criterion

FC is a parametric criterion that used linear projection to convert high dimensional data into one single dimension. This technique basically minimized the variance within a group and attempts to maximize the mean between difference groups. FC is defined as given the formula below:

$$FC = \frac{|m_1 - m_2|^2}{s_1 - s_2} \quad (2)$$

where m1 and m2 are the mean of cancerous and non- cancerous class and s is the variance.

2.2.3 Information gain

IG is a popular gene selection technique used to obtain genes that can differentiate two different groups. The underlying concept of this technique is from the information theory. It use an entropy function to calculate the weight of a feature to partition the tissue samples. Formula of IG is as illustrated below:

$$IG\ (F,S) = \sum_{v=values\ F} \frac{|S_v|}{|S|} (Entropy\ (S) - Entropy\ (S_v)) \quad (3)$$

where F is possible genes and S is proportion of classes

2.2.4 T-test

T-test is a statistical hypothesis that measures the differences between two classes when normal distribution of these classes is unknown. T-test is measured as shows below:

$$T - test = \frac{|\mu i^+ - \mu i^-|}{\sqrt{\frac{(\sigma i^+)^2}{n^+} + \frac{(\sigma i^-)^2}{n^-}}} \quad (4)$$

Where μ is the mean and δ is standard deviation for two difference classes, n + = cancerous samples and n- = non-cancerous samples.

2.3 Validation phase

Receiver Operating Characteristics (ROC) curve is used in this study to calculate and determine the performance of filter-based gene selection techniques. In addition, various other metrics are used to measure ability of these techniques to separate the tissues sample into different groups. Formula one of these metrics are given below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

Where TP is true positive; TN is true negative; FP is false positive; and FN is false negative

Moreover, the proposed techniques are also validated by using 10-fold cross validation

3. Experimental results and discussion

In this study, comparative analysis of four filter-based gene selection techniques, namely SNR, FC, IG and T-test are conducted to determine the discriminative power of SVM and C4.5 classifier. At the beginning, prior the downstream analyses of gene selection method, the selected gene expression datasets have been analyzed using heat map. A heat map is a graphical representation of data where the individual values contained in a matrix. Genes are portrait in rows while samples are represented in columns. The values of gene expression, Xij, on the other hands are visualized in a color form, where green color denotes to down regulated genes (under expressed), and red color signifies upregulated genes (over expressed). The obtained publicly genes expression dataset have been analyzed using heat map to determine the distribution of gene expression. This analysis is executed using R function in a biocLite R package.

Figure 3 shows the performance of gene selection techniques. The first row illustrated results obtained when was run with SVM and second row is for C4.5. Detail results are presented in Table 2 and Table 3.

Table 2: The accuracy results of SVM classifier on ten-fold cross validation

Dataset/ Number of genes	Original	500 genes				200 genes				100 genes				50 genes			
		SNR	FC	IG	T	SNR	FC	IG	T	SNR	FC	IG	T	SNR	FC	IG	T
Breast cancer	.63	.74	.60	.78	.45	.76	.58	.80	.50	.70	.57	.75	.61	.68	.57	.70	.60
Colon cancer	.80	.87	.82	.88	.80	.87	.80	.90	.82	.85	.81	.87	.80	.82	.80	.85	.82
Lung cancer	.83	.87	.60	.87	.50	.89	.63	.85	.51	.91	.65	.84	.50	.89	.68	.80	.50

Table 3: The accuracy results of C4.5 classifier on ten-fold cross validation

Dataset/ Number of genes	Original	500 genes				200 genes				100 genes				50 genes			
		SNR	FC	IG	T	SNR	FC	IG	T	SNR	FC	IG	T	SNR	FC	IG	T
Breast cancer	.57	.60	.60	.78	.40	.65	.58	.80	.40	.57	.57	.75	.41	.60	.57	.70	.40
Colon cancer	.70	.72	.65	.88	.70	.70	.70	.90	.80	.70	.81	.87	.80	.70	.80	.85	.82
Lung cancer	.80	.70	.75	.87	.60	.70	.63	.85	.41	.70	.65	.84	.50	.70	.68	.80	.50

Experimental results in Table 2 and Table 3 have shown that gene selection is an important and crucial task in enhancing the accuracy of classifiers. The results have revealed that classifier performance could be further increased in comparison to the used of original datasets, for both classifier; SVM and C4.5. The comparative analysis has shown that SNR and IG techniques work better compare to other filter-based gene selection techniques when run with SVM, with mean accuracy of 83% and 84%, respectively.

This study also has discovered that SNR and IG are more suitable to be used with SVM while IG fit perfectly for C4.5. In a colon dataset, SVM has achieved a specificity of 86% with SNR while and 80% for IG. In contract, C4.5 has obtained a specificity of 78% for IG on the identical dataset. These results indicate that SVM performed slightly better with IG pre-processed data compare to C4.5 on the same dataset. Furthermore, it has been discovered that these filter-based gene selection techniques, SNR, IG, FC and T-test performed better in colon dataset in comparison to

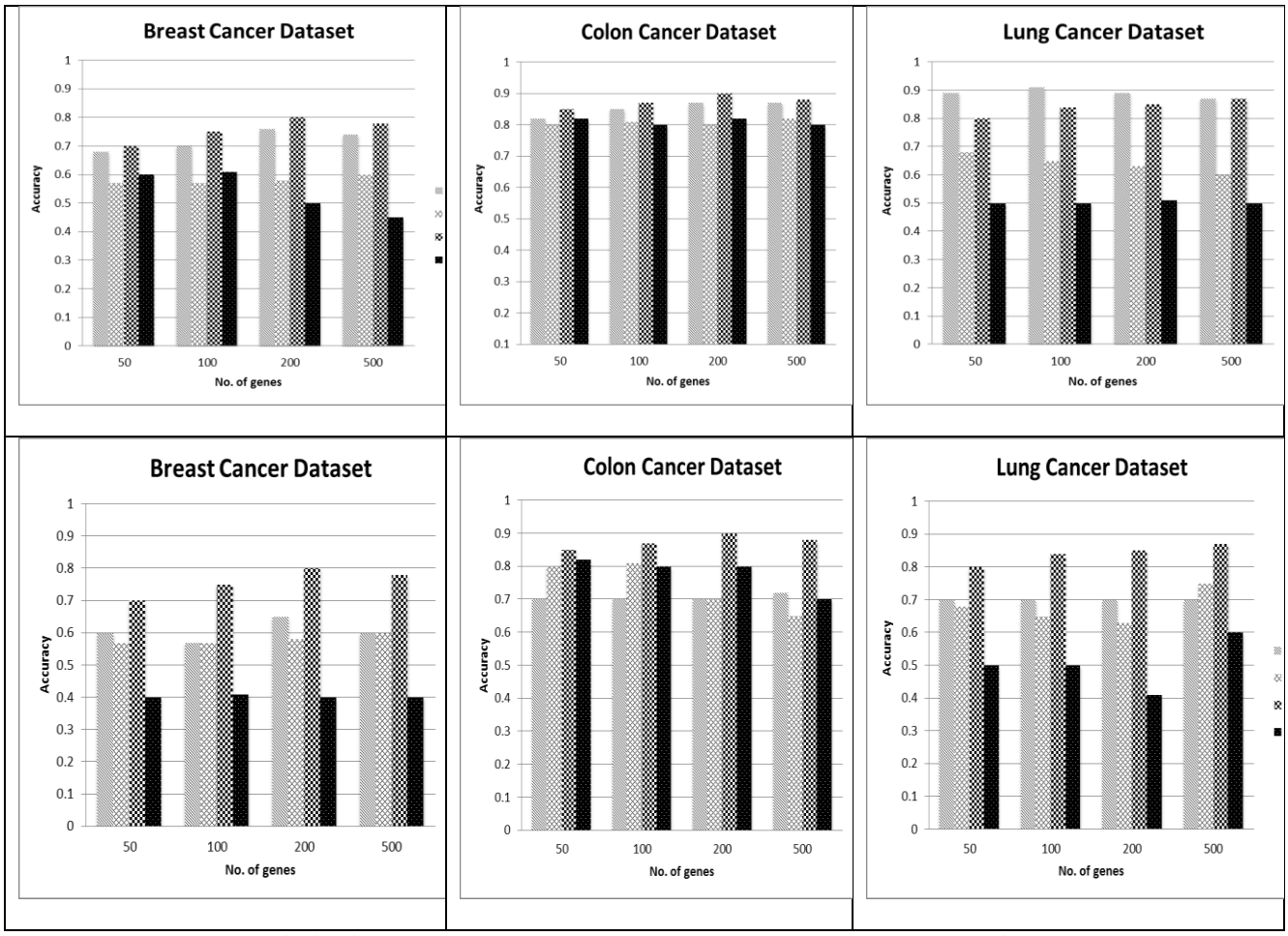

Figure 3: Experimental results for four filter-based gene selection technique, SNR, FC, IG and T-test tested on SVM (first row) and C4.5 (second row)

breast cancer or lung dataset. It is most probably due to the small number of genes in relative to other datasets which have been tested. Hence, other subsequent analysis is required to address the high dimensional gene expression profiles.

4. Conclusion

DNA microarray is the breakthrough technology in domain of biology. However, due to the nature of this microarray data, it is common that a large number of genes are not informative for classification because these genes can be either irrelevant or redundant. For such a reason, gene selection methods become apparently needed. In this study, we aim to investigate and compare the effectiveness of these four popular filter gene selection techniques such as SNR, FC, IG and t-Test in selecting informative genes that can distinguish cancer and normal tissues. Our experimental results have shown that gene selection techniques are equally importance in classifying tissues samples besides having an appropriate classifier. Hence, by go through gene selection process the classification accuracy can be further improved.

Acknowledgement

This work was supported in part by University Grant under Grant Nos. 12866.

References

[1] K. Kourou, & D. I. Fotiadis, (2015). Computational modelling in cancer: Methods and applications. Biomedical Data Journal, 1(1), 15–25.

[2] R. Hu, (2011). Medical data mining based on decision tree algorithm. Computer and Information Science, 4(5), 14–19.

[3] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A.M Adrian, W..-C. Cheng, … K.-S. Chang, (2014). Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics, 15(1), 49.

[4] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, … A. Nowé, (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106–19. doi:10.1109/TCBB.2012.33

[5] F.K. Ahmad, S. Deris, N.H. Othman, and N.M. Norwawi, (2009). A review of feature selection techniques via gene expression profiles. IEEE International Symposium on Information Technology (ITSim), 2008, Kuala Lumpur, Malaysia.

[6] O.F. Huey, N. Mustapha & N. Sulaiman, (2011). Integrative gene selection for classification of microarray data. Computer and Information Science, 4(2), 55–63.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.

[8] L. Li, and H. Li, (2004). Dimension reduction methods for microarrays with application to censored survival data. Briefings in Bioinformatics, 20(18), 3406-3412.

[9] O. Gevaert, F.D. Smet, D. Timmerman, Y. Moreau, and B.D Moor, (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22(14), e184–e190.

[10] Y. Saeys, I. Inza, and P. Larranaga, (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-251.