Identification of Genes of Melanoma Skin Cancer Using Support Vector (SVM)-Based Classification

***Nurhafizah Mohd Nazari*[1], *Zuraini Ali Shah**[2]**

Department of Software Engineering, Faculty of Computing,
Universiti Teknologi Malaysia,
81310 Johor Bharu, Johor, Malaysia

[1]*nurhafizah.nazari@gmail.com*, [2]*aszuraini@utm.my*

Abstract

Melanoma skin cancer is the serious type of cancer. Genetic modification can lead to melanoma. This disease can be cured if diagnosed early. In cancer classification, the main problem is the way to handle large data of cancer. Large data can lead to unclean data which means the data contains meaningless data, redundant data and much more. Therefore, the classification process may not too accurate. In this research, the problem in classification is to identify the genes of melanoma with the high dimensional datasets. When the datasets too big, it is hard to understand the causal mechanism of the disease. To overcome the problem, many computer approaches for classification have been proposed in the previous literature. Therefore, the purpose of this research is to identify genes of melanoma skin cancer using Support Vector Machine (SVM)-based as the classifier. The different kernel function is used to identify which kernel gives the best performance of classification toward the melanoma datasets. To identify the specific genes related to melanoma, data sets of gene expression related to melanoma are used in this research. Before classification method is conducted, pre-processed of genes expression data is done first by using features selection techniques to identify the most significant gene and to reduce the dimensional of the datasets. Next, the classification process for melanoma is conducted by using SVM-based with a different kernel in this research. This research found that linear kernel and polynomial kernel obtain the best result in classification towards melanoma datasets.

Keywords: Big data, classification, SVM

1.0 Introduction

Cancer is a disease that involved the abnormal cell that can spread to other parts of tissues. There are many types of cancer such as skin cancer, leukemia, lung cancer, breast cancer and much more. The cancer disease can be diagnosed if early detection is made.

Skin cancer is incredibly common among the humans. Skin cancer happened when abnormal cells grow in the skin in a dangerous manner. Sometimes, they are growing in the lymph nodes, or other organs in our body. Researchers found that the skin cancer rising fast in United Kingdom (U.K). The main factor of skin cancer is causes of sun exposure. Besides, skin cancer also affected from the genetic problem. There are three common types of skin cancer which are basal cell skin cancer, squamous cell skin cancer and two major groups of skin cancer, melanoma skin cancer, and non-melanoma skin cancer.

In previous literature, the main problem in classifying a cancer is when using the gene expression data, (Lu and Han, 2003). Gene expression data is a big and high dimensional type of data. Hence, analysis of gene expression is difficult to conduct because of the big data that contain noise, redundant data, unrelated information of features and much more. Therefore, to overcome those problems a method such as a feature selection is applied, (Lu and Han, 2003).

In this research, a computational approach is used for classifying the unknown tissues sample. A computational method such as features selection technique is used for feature selection. Support Vector Machine algorithm (SVM) is used as the classifier. Support Vector Machine is a learning and supervised machine that is used for classification and regression. Most of the previous literature mentioned that SVM is the best algorithm to use for classification. Other than that, most of the previous literature proved that SVM provides high accuracy for the classification technique.

Most of the classification techniques for classifying the cancers by using computational-based have been developed before. However, there are still a lot of improvement to make for a better result on classifying the cancers. At the same time, the improvement that researchers make for the classifier also lead to improvised the process of diagnosing the cancerous illness. In order to perform the better and more accurate for the classifying system, a lot of challenges should be handled. In a gene-based classification task, the common challenges that need to handle is the high dimensional of microarray data. The main problem is to reduce the dimensionality and to overcome the risk of overfitting. Sometimes the information of the features in the data is hard to understand or low quality of data. To handle the classification problem, many computational methods for classification is proposed .

Three objectives have been identified for this research. The first objective is to study the domain of skin cancer and the computational approach that related with Support Vector Machine (SVM)-based. Second, to identify the genes of melanoma skin cancer using feature selection and classification techniques. And the last objective is to evaluate the performance of the accuracy of SVM classification on the genes of melanoma skin cancer.

2.0 Methodology

Gene expression of melanoma skin cancer (GSE3189) is obtained from GEO database. The platform of the data used is Affymetrix Human Genome U133A Array. GSE3189 contains three types of classes which are normal skin, nevi skin, and melanoma. This set of data contains 70 samples. 7 of them are normal skin, 18 of them are nevi skin and the other 45 are melanoma samples. From the data collected, pre-processed of data is done by using normalization and features selection techniques to reduce the dimensional and noise of the data. Next, the pre-processed data is used for classification process. Classification process is done by using WEKA tool with different types of kernels. The results for classification of different kernel is are recorded. From the classification results, the performance measurement of the classifier is calculates. The performance measurement of the classifier is recorded based on the precision value, recall value and ROC value.

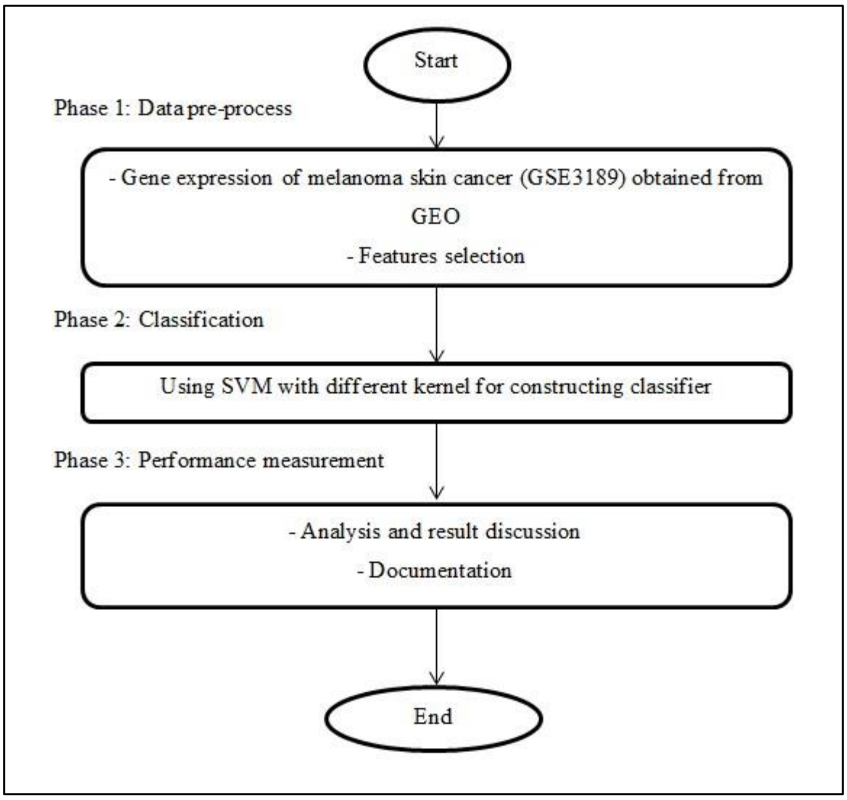
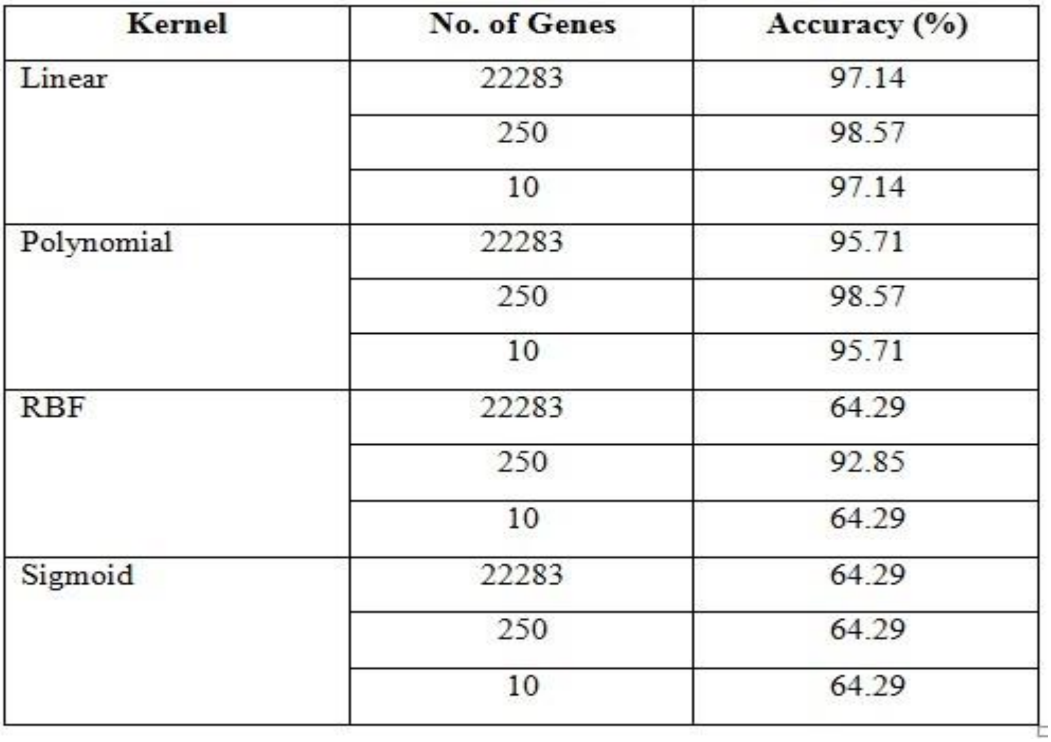

Start
Phase 1: Data pre-process
- Gene expression of melanoma skin cancer (GSE3189) obtained from GEO
- Features selection
Phase 2: Classification
Using SVM with different kernel for constructing classifier
Phase 3: Performance measurement
- Analysis and result discussion
- Documentation
End

Figure 1 The overall flow of research methodology

3.0 Result

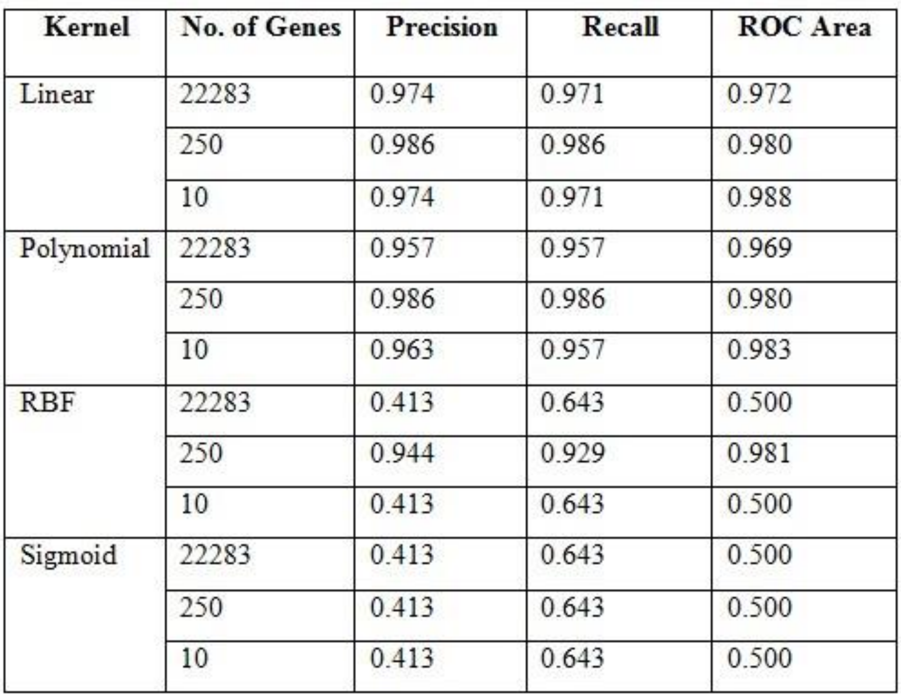
Below are the result for classification using SVM with different types of kernels which are linear kernel, polynomial kernel, radius basis function kernel (RBF) and sigmoid kernel. The classification result is display on Table 1. The result for classifier performance also is display on Table 2.

Table 1: Result of classification using SVM with different kernel



Kernel	No. of Genes	Accuracy (%)
Linear	22283	97.14
	250	98.57
	10	97.14
Polynomial	22283	95.71
	250	98.57
	10	95.71
RBF	22283	64.29
	250	92.85
	10	64.29
Sigmoid	22283	64.29
	250	64.29
	10	64.29

Table 2: Classifier performance



Kernel	No. of Genes	Precision	Recall	ROC Area
Linear	22283	0.974	0.971	0.972
	250	0.986	0.986	0.980
	10	0.974	0.971	0.988
Polynomial	22283	0.957	0.957	0.969
	250	0.986	0.986	0.980
	10	0.963	0.957	0.983
RBF	22283	0.413	0.643	0.500
	250	0.944	0.929	0.981
	10	0.413	0.643	0.500
Sigmoid	22283	0.413	0.643	0.500
	250	0.413	0.643	0.500
	10	0.413	0.643	0.500

4.0 Discussion

The accuracy for three different values of genes are quite high which is above than 90% accuracy. Because of the linear kernel is the simplest kernel for SVM, it can divide the dataset into training and testing accurately. Which mean, the linear kernel is good to use in SVM for classification. Other than the linear kernel, the polynomial kernel also gives a high accuracy of classification. The accuracy of classification obtain from different numbers of genes are greater than 90%. This also means that polynomial kernel is good to use in classification. Comparing between linear kernel and polynomial kernel, the linear kernel gives the best result of classification.

But not too RBF kernel and sigmoid kernel. Those kernels obtain lower accuracy of classification than the linear kernel and polynomial kernel. But an odd result happened to the 250 genes during the classification. It obtains high result than other numbers of genes which are 92.85%. This kind of result may due to the value of differential expressed genes and its behaviour. Also, may due to the optimum number of genes that RBF can work accurately. The accuracy for the sigmoid kernel for different numbers of genes are same. It may cause due to the optimal parameter value used by sigmoid function during the classification.

Based on the classification result, the linear kernel is the best kernel that can be used in SVM for classification by using melanoma datasets.

Based on Table 2 above, the best performance obtain is the linear kernel. The precision, recall and ROC area value are almost 1. Which mean the performance of classifier using linear kernel is almost accurate. Compare with the polynomial kernel, the differences of the performance with the linear kernel not too far. It means that polynomial kernel in SVM also gives the most accurate of the classifier. Meanwhile, RBF kernel and sigmoid kernel give the lower performance than the linear kernel and polynomial kernel. Based on the overall performance measurement result, the best kernel that can be used in this research is linear kernel towards the melanoma datasets.

5.0 Conclusion

In short, this research focusses on to identify the result of classification using SVM with different kernel and to identify which kernel is the best. Classification result shows linear kernel gives the high accuracy of classification towards the melanoma data which is higher than 90%. From the classification process, 10 best genes are identified related to melanoma skin cancer. To evaluate the SVM performance, performance of measurement for is constructed. The results of performance shows that, SVM with linear kernel gives the best performance than other kernels. Therefore, proved that SVM with linear kernel can classify the selected genes accurately.

References

Jain, Y. K., & Jain, M. (2012). Comparison between different classification methods with application to skin cancer. skin, 1, 3.

Liu, D., Liu, X., & Xing, M. (2014). Activities of multiple cancer-related pathways are associated with BRAF mutation and predict the resistance to BRAF/MEK inhibitors in melanoma cells. Cell Cycle, 13(2), 208-219.

Liu, W. (2015). An Integrated Bioinformatics Approach for the Identification of Melanoma Associated Biomarker Genes. A Ranking and Stratification Approach as a New Meta Analysis Methodology for the Detection of Robust Gene Biomarker Signatures of Cancers (Doctoral dissertation, University of Bradford).

Singh, P. K., & Karthikeyan, S. (2012, May). Combining GRNN and SVM Using Receiver Operating Characteristics (ROC) for Improved Classification of Non Coding RNA. In Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on (pp. 115-118). IEEE.

Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information Systems, 28(4), 243-268.