Evaluate the Performance of SVM Kernel Functions for Multiclass Cancer Classification

***Noramalina Mohd Hatta*[1]*, Zuraini Ali Shah**[2]*

*Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia
[1]are.mallyna93@gmail.com, [2]aszuraini@utm.my*

\

Abstract

Multiclass cancer classification is basically one of the challenging fields in machine learning which a fast growing technology that use human behaviour as examples. Supervised classification such Support Vector Machine (SVM) has been used to classify the dataset on classification by its own function and merely known as kernel function. Kernel function has stated to have a problem especially in selecting their best kernels based on a specific datasets and tasks. Besides, there is an issue stated that the kernels function have a high impossibility to distribute the data in straight line. Here, three basic kernel functions was used and tested with selected dataset and they are linear kernel, polynomial kernel and Radial Basis Function (RBF) kernel function. The three kernels were tested by different dataset to gain the accuracy. For a comparison, this study conducting a test by with and without feature selection in SVM classification kernel function since both tests will give different result and thus give a big meaning to the study.

Keywords: Multiclass Cancer Classification, SVM, Kernel Function, Machine learning

1.0 Introduction

In bioinformatics fields, genes identifications are responsible for classifying existing disease samples of two or more of its variants. As previous study had been done and solved involving supervised learning methods such as k-nearest neighbour (KNN), weighted voting approach, support vector machine (SVM), linear discriminant analysis (LDA), artificial neural networks (ANN) and Random forest. Besides, in cancer classification using microarray data, an increasing number of studies have successfully demonstrated the effectiveness of state-of-the-art supervised machine learning methods such as Support Vector Machines (SVMs). SVMs are defined as powerful classification machine learning based on the variety of regularization technique (Niijima and Kuhara, 2005). The SVMs were built to learn a function that generates output based on input and for the next new output can be easily generated since the old function has learned from the previous case.

The aim of this project is to analyze the technique of Support Vector Machine (SVM) kernel function of linear kernel, polynomial and RBF kernel function that related to multiclass classification cancer. Next is to analyze the implementation of SVM classification to get the best kernel function by accuracy and computational time on multiclass cancer. Lastly, to evaluate the implementation of SVM kernel functions for multiclass cancer classification by obtaining the accuracy and time taken.

2.0 Background of The Study

For gene expression data, there are several issues that need to be alert. The main difficulties for solving the result of optimization problem is the gene expression data is in a high dimensional with small but significant uncertainty in the original labelling's and the noise of the experimental and measurement process and the intrinsic biological variation from specimen to specimen is difficult in enhancing optimization (Ramaswamy et al., 2001). Next is gene expression data is tends to redundant, bias and confusing problem which make a classification more difficult and causing slow performance and used too much time. Lastly, for this such problems can also be posed as optimization problems of minimizing gene subset size to achieve reliable and accurate classification (Deb and Reddy, 2003). Previous research revealed that a multiclass cancer classification can be classified by SVM and among well-established and popular techniques for classification of microarray gene expression data, SVM achieve the best classification performance (George and Raj, 2011) because of the output was constructed in hyperplane within infinite dimensional space which are linear and nonlinear.

Moreover, SVM depends on the standard of Structure Risk Minimization by taking into record of the likelihood of misclassifying yet to be seen designs for an altered however obscure likelihood conveyance of information. It utilizes a direct isolating hyperplane to make a classifier, yet it is difficult to partition a few issues in the first info space directly. Be that as it may, it can effectively change the first info space into a high dimensional component space nonlinearly, where it is minor to locate an ideal direct isolating hyperplane. The standard Support vector machine algorithm is prompts a quadratic enhancement issue with bound requirements and one direct fairness limitation. Be that as it may, when the datasets are substantial with extensive number of information focuses, the quadratic programming solvers turn out to be exceptionally troublesome, on the grounds that their time necessities. Furthermore, memory is very reliant on the span of the preparation datasets.

Thus, this research focuses on evaluate the performance of SVM kernel function in finding the best function among linear kernel, polynomial and radial basis function (RBF) kernel.

3.0 Methodology

In this research, the multiclass cancer classification is the main focus. Firstly, identification of problem statement and current methods of this research was identified. The objectives of the study, aim and scope were centralized according to all problems listed.

Following after that is research planning such as methods and operational framework and the data set and library searching.

Secondly, on the second phase, the preprocessing data set is obtained by collecting all sets of data referencing. All the data set has gone through preprocessing step and has undergo normalization of 0 and 1. Which rescaling the data on training set for maximize the optimization for classification. The datasets are DLBCL (Shipp et al.,2002), brain tumor (Pomeroy et al., 2002) and 9_Tumors (Stuanton, 2001). Also, in the datasets obtain all the data is converted to MATLAB format (.mat). The dataset and the classes species is separated to use in classification.

Third, at the phase three, with and without feature selection and implementation of SVM kernel function was conducted. The classifiers with different kernel functions namely as linear kernel, polynomial kernel and RBF kernel were tested for every datasets in classification after obtaining subset from feature selection and without feature selection the data immediately use in SVM classification. The rank of genes in feature selection was tested according to a certain subsets and for without selection, the actual values of samples against gene were straight away used in SVM classification while using the kernels function.

The final phase was testing and analyzing. The classification was happened due to the data set is put onto SVM classifier to do their work and at the same time, the process was happened to optimize the SVM parameter and get the accuracy of the result. Here, the evaluation on data sets, analyzing the data and comparing the accuracy result of classification to get the best kernel function will be conducted.

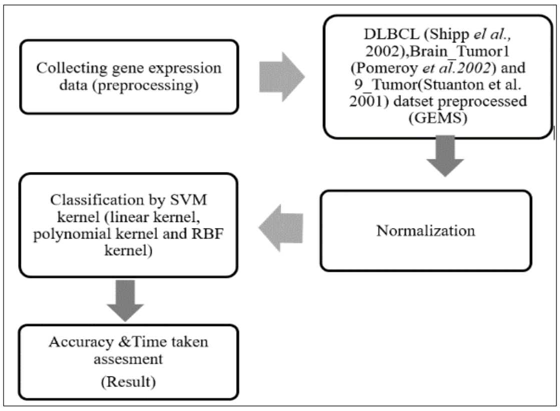
Collecting gene expression data (preprocessing)
DLBCL (Shipp *el al.,* 2002),Brain_Tumor1 (Pomeroy *et al.2002*) and 9_Tumor(Stuanton et al. 2001) datset preprocessed (GEMS)
Normalization
Classification by SVM kernel (linear kernel, polynomial kernel and RBF kernel)
Accuracy &Time taken assesment
(Result)

Figure 1 The flow chart of classification of SVM kernel function without feature selection

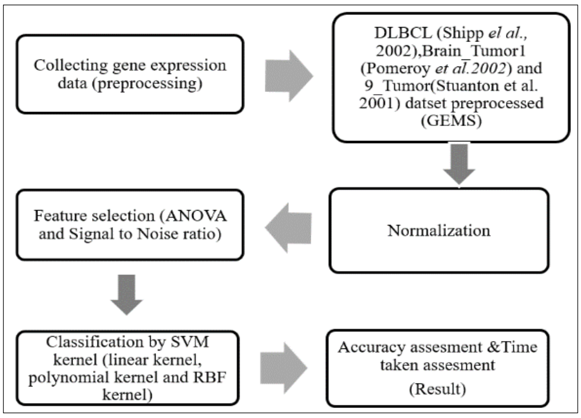
Collecting gene expression data (preprocessing)
DLBCL (Shipp *el al.*, 2002),Brain_Tumor1 (Pomeroy *et al.2002*) and 9_Tumor(Stuanton et al. 2001) datset preprocessed (GEMS)
Normalization
Feature selection (ANOVA and Signal to Noise ratio)
Classification by SVM kernel (linear kernel, polynomial kernel and RBF kernel)
Accuracy assesment &Time taken assesment
(Result)

Figure 2 The flow chart of classification of SVM kernel function with feature selection

4.0 Results

In Table 1, the performance analysis was executed without feature selection.

Table 1 Performance analysis of different kernel of SVM

Dataset	Kernel Function			Kernel Function		
	Accuracy (%)			Time taken to build the model (s)		
	Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	**97.40**	24.68	75.32	8.349	4.062	**3.403**
Brain	**94.44**	36.67	66.67	7.596	4.184	**3.427**
9 Tumor	**85.00**	15.00	**85.00**	3.781	4.159	**3.065**

In Table 2, shows the comparison of different kernel function accuracy result, time taken to build the model as the minor analysis against dataset after ANOVA test were done and classification process was executed.

Table 2 Performance analysis of ANOVA test with different kernel of SVM classification

Dataset/No. of Genes		Kernel Function Accuracy (%)			Kernel Function Time taken (s)		
		Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	10	**90.91**	79.22	80.52	6.006	7.467	**1.854**
	100	**81.82**	79.22	75.32	7.524	6.174	**1.818**
Brain	10	**77.78**	68.89	75.56	4.501	6.986	**1.591**
	100	**88.33**	78.89	66.67	8.491	4.916	**1.92**
9 Tumors	10	**88.33**	86.67	86.67	5.366	6.238	**2.426**
	100	**88.33**	76.67	85.00	5.043	3.531	**1.903**

For Table 3, the result shown is the comparison of kernel function accuracy result, time taken to build the model as the minor analysis against dataset after Signal to noise ratio test were done and classification process was executed.

Table 3 Performance analysis of Signal to noise ratio test with different kernel of SVM classification

Dataset/No. of Genes		Kernel Function Accuracy (%)			Kernel Function Time taken (s)		
		Linear	Polynomial	RBF	Linear	Polynomial	RBF
DLBCL	10	90.91	75.32	**92.21**	6.038	6.789	**1.854**
	100	**96.10**	76.62	75.32	7.082	6.642	**1.818**
Brain	10	**78.89**	75.56	73.33	6.130	7.314	**1.591**
	100	**86.67**	77.78	66.67	8.898	3.423	**1.92**
9 Tumors	10	**90.00**	88.33	86.67	4.795	4.205	**1.967**
	100	**90.00**	81.67	85.00	4.607	3.084	**2.801**

5.0 Discussion

Based on the result from previous section, the high accuracy obtained for the whole dataset by without feature selection is linear kernel function. Whilst, the dataset that have high accuracy of 90 percent and above is containing the informative genes which makes the accuracy higher. With this, it shows that eventhough linear kernel was set a record as best classification of two classes, it also shows that linear kernel function is good with multiclass classification. With the domination of accuracy number, shows that the function gives a good lesson to the testing dataset for multiclass problem without feature selection. Also, with the linear kernel function, it shows that these dataset are suited more with this kernel function compared to other kernel function in classifying the data. However, the time taken to build the model was evaluated but its shows that the high accuracy need more time to compute the classification. And thus, the linear kernel function was selected to be the best kernel function.

Next, the high accuracy obtained for the whole dataset for both test by feature selection are linear kernel. But for a Signal to noise ratio feature selection, one dataset give different number accuracy which lowering the domination of linear kernel function for the whole feature selection test classification result. The data that give one difference is DLBCL dataset with 10 samples. Whilst, the dataset that have high accuracy of 90 percent and above is actually containing the informative genes and thus the noisy data was lessen. For liner kernel function, with the domination accuracy number, shows that the function give a good lesson to the testing dataset for multiclass problem. Also, with the linear kernel function, it shows that these dataset are suit more with this kernel function compared to other kernel function in classifying the data. In addition, brain dataset have low number of informative gene when the selection of genes was made. The dataset shows clearly in the result that when the genes tested have high number, the accuracy obtained will be also high. However, the time taken to build the model was evaluated but its shows that the high accuracy need more time to compute the classification. And thus, the linear kernel function was selected to be the best kernel function.

6.0 Conclusion

Cancer classification is one of the challenging tasks especially using a machine learning named as Support Vector Machine (SVM). SVM is known to be good in classifying binary classes which it made difficult to deal with. Therefore, there are lots of researchers have proposed numerous attempts in classifying multiclass cancer using SVM. The main goal was to classify the data to get the best accuracy by comparing SVM kernel function by a help of with and without feature selection before the classification process. MATLAB 2013b was used in this study to point out the classification by performing accuracy of DLBCL, brain and 9 tumors dataset.

This research was conducted in to help solving the problem statement of this study as well as comparing the accuracy by with and without feature selection. In aiming to get which is the best kernel function for all the dataset tested, the experiment was performed and the results were recorded. The kernel function used for comparison is linear kernel, polynomial and radial basis function (RBF) kernel function. Also, these kernel functions give goods result in some datasets.

References

Deb K., and Reddy A. R. (2003). Reliable classification of two-class cancer data using evolutionary algorithms. BioSystems, 72(1), 111-129.

George G., and Raj V. C. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile.arXiv preprint arXiv:1109.1062.

Niijima S., & Kuhara S. (2005). Effective nearest neighbor methods for multiclass cancer classification using microarray data. In Proceedings of the 16th International Conference on Genome Informatics P (Vol. 51).

Pomeroy S. L., Tamayo P., Gaasenbeek M., Sturla L. M., Angelo M., McLaughlin M. E.,and Golub T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature,415(6870), 436-442.

Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., and Golub T. R. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences,98(26), 15149-15154.

Shipp M. A., Ross K. N., Tamayo P., Weng A. P., Kutok J. L., Aguiar R. C., and Ray T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine, 8(1), 68-74.

Staunton J. E., Slonim D. K., Coller H. A., Tamayo P., Angelo M. J., Park J., and Mesirov J. P. (2001). Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences,98(19), 10787-10792.