World Journal of Research and Review (WJRR)
ISSN: 2455-3956, Volume-7, Issue-4, October 2018, Pages 25-27.

Comparative Evaluation of Filter and Wrapper Based Approach for Micro Array Cancer Classification

Akinrotimi Akinyemi Omololu, Mabayoje Modinat Abolore

***Abstract*— The microarray classification technique is capable of tracking the expression levels of thousands of genes simultaneously. The high dimensional feature vectors of microarray, impose a high dimensional cost, as well as the risk of over fitting during classification. Thus, it is necessary to reduce the dimension, through the use of an efficient method such as; feature selection. In this paper, a different technique, based on classification of micro array cancer dataset, including the two basic approaches of feature selection: the filter and wrapper techniques have been employed.**

***Index Terms* — Classification, Filter Approach, Wrapper Approach, Support Vector Machine, Kernel, Microarray.**

I. INTRODUCTION

The microarray technology is one of the most efficient tools in recent times for finding out the levels of expressions of genes, in a concurrent manner. The classification of the samples of patients based on gene expression in order to diagnose and treat diseases stemmed from the hypothesis that; virtually all human disease are caused by certain gene expressions.

The amount of data that is needed in carrying out a cogent analysis in machine learning increases as the data increases. Bellman referred to this phenomenon as the "curse of dimensionality" when considering problems in dynamic optimization [1]. One of the most wildly used methods in solving the problem of "curse of dimensionality" is searching for a projection of data onto smaller features which is capable of preserving information as much as possible. Each data point (sample) can have up to 450,000 variables (gene probes) and processing a large number of data points involves high computational cost [2]. However, the difficulty of getting enough of meaningful data, with increasing dimensionality of a data set results in mounting difficulty in statistically proving the result. Larger data sets are often susceptible to over fitting. This can lead to errors in the classification of the data when the over fitted model takes small but frequent changes for important variance in the data or when there are noisy variables within the dataset. Noise in a dataset is defined as "the error in the variance of a measured variable" which can result from errors in measurements or natural variation [3]. In machine learning, noisy data should be reduced as much as possible order to reduce or totally evade complexity, in inferred models, thereby improving the efficacy of the algorithm.

Cost of computation increases with increasing dimensionality, however, to assuage this problem, the number of features in the data set is often reduced. To reduce the number of features in a data set, the two common techniques used, are: feature subset selection and feature extraction. Considering the efficacy of the two features, this research work carried out a comparative evaluation of filter and wrapper approaches with multiple SVM kernels for the classification of micro array cancer dataset using the Colon tumor dataset.

II. LITERATURE REVIEW

Quinlan [5] proposed a classification algorithm called ID3, which introduces the concept of information gain. Information gain is a measure based method, which is usually used to select best split attributes in decision tree classifiers. The measure indicates to what extent the entire data's entropy is reduced, and identifies the value of each specific attribute. Each feature basis obtains an information gain value, the amount of which is used to decide whether the feature is selected or deleted. [6] used DNA microarrays to conduct a systematic characterization of gene expression in B-cell malignancies. The expression patterns of patients with diffuse large B-cell lymphoma (DLBCL) were studied. Hierarchical clustering with average linkage search was used on the gene expression patterns of 88 biological samples to identify two previously unidentified molecularly distinct forms of DLBCL (germinal centre B-like DLBCL and in vitro activated peripheral blood B-like DLBCL) , which had gene expression patterns indicative of different stages of B-cell differentiation. They equally demonstrated that patients with the two sub-groups of tumor are susceptible to different clinical outcomes. [7] used an hierarchical clustering method on expression patterns of lung cancer patients to identify patients with various kind of this cancer type that are characterized by different prognostic outcomes. Yu and Liu [8] proposed a fast correlation-based filter algorithm (FCBF) which used correlation measure to obtain relevant genes and to remove redundancy. Ding and Peng [9] have used mutual information for gene selection that has maximum relevance with minimal redundancy by solving a simple two-objective optimization. Xing et al. [10] proposed a hybrid of filter and wrapper approaches to feature selection.

Akinrotimi Akinyemi Omololu, Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria, P.M.B 1515, Ilorin, Kwara State, Nigeria

Mabayoje Modinat Abolore, Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria, P.M.B 1515, Ilorin, Kwara State, Nigeria

III. PROPOSED SYSTEM

A. Technique Procedure

The Colon tumor dataset was experimented upon which consists of 62 microarray experiments collected from colon-cancer patients with 2000 gene expression levels. Among them, 40 tumor biopsies are from tumors and 22 (normal) biopsies are from healthy parts of the colons of the same patients.

(b) Grouping of Data Set/Normalization

Normalization is necessary when comparing the results of different microarray dataset, because it removes variation caused by the manufacture, preparation or experimental handling of the datasets.

(c) Feature Selection using the Filter Approach

The input variables are the risk elements or factors which put a lady at a higher risk of getting cervical cancer. The inputs are:

(i) Correlation-based feature selection (CFS):

CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [8].The equation is shown below:

$$\mathrm{CFS}_S = \frac{k\bar{r}_{\mathrm{cf}}}{\sqrt{k + k(k-1)r_{\mathrm{ff}}}} \qquad \text{.....Equ. 3.1}$$

where, CFS $_S$ is the score of a feature subset S containing k features, $r$cf is the average feature to class correlation (f $\in S$), and $r$ff is the average feature to feature correlation. The distinction between normal filter algorithms and CFS is that: while normal filters provide scores for each feature independently, CFS presents a heuristic "merit" of a feature subset and reports the best subset it finds.

(ii) SVM Wrapper base

For the wrapper based approach the dataset was wrapped around the SVM classifier.

B. Classification

In SVMS training data is analyzed and used for classification. SVM is used in analyzing the data by imputing a set of training data with positive and negative classes into it. Thereafter, the SVM creates a decision boundary between the classes and picks the most relevant classes that are relevant in the decision process. Provided the data is linearly separable, it will always be possible to construct a linear boundary. Where the data is linearly inseparable, SVMs will make use of kernels which projects the data into a higher dimensional feature space.

C. Performance Evaluation

The classification of the dataset was done, using the filter and wrapper methods and the support vector machine. The statistical result obtained from this machine learning algorithm is used to predict the best method for feature selection.

D. System Framework

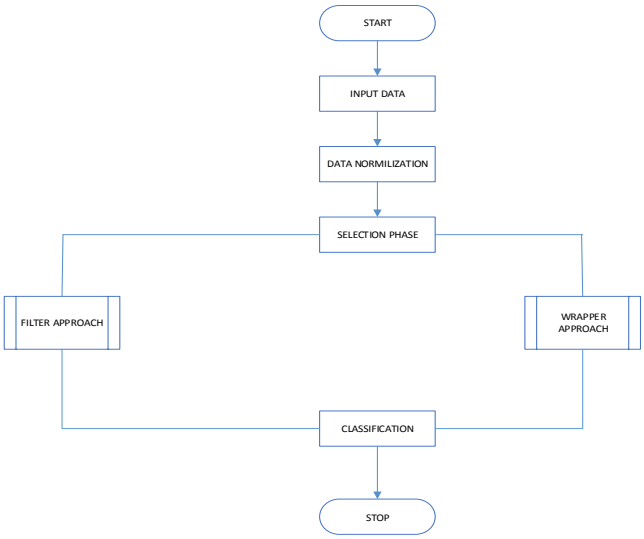

Figure 3.1: System Architecture

IV. SYSTEM IMPLEMENTATION AND RESULTS

The Colon tumor dataset was experimented upon which consists of 62 microarray experiments collected from colon-cancer patients with 2000 gene expression levels. Among them, 40 tumor biopsies are from tumors and 22 (normal) biopsies are from healthy parts of the colons of the same patients. The filter base approach follows suit, the use of the correlation feature selection, to select relevant and discrete data from the large data set which gave a reduction from 2000 attributes to 27 attributes which was passed to the classifier using support vector machine. A case Study of three support vector kernels was duly examined and processed for the classification task namely the Normalized Poly kernel, Poly Kernel and Radial Bias Function Kernel.

The Wrapper based Approach was also examined and experimented upon with the same dataset by using a subset evaluator so as to create all possible subsets form the feature vector, after which our classifier algorithm was used to induce classifiers from the features in each subset by considering the subset of features whose classification performs the best. After feature selection, the selected feature subsets were evaluated using Sequential Minimization Optimization Technique (SMOT). As for the classification algorithms, SVM are generated by repeating the 10-fold cross-validation method ten times. The classification models were built and tested by using the Support vector machine algorithms using three varying kernels: Radial Bias Function Kernel, Polynomial kernel and the Normalized Polynomial kernel. The Best First search technique was invoked to find a subset an evaluator will use.

Comparative Analysis of Wrapper and Filter Based Approach Used

Based on the Colon tumor dataset, both the filter based and wrapper based approach are being evaluated on the Colon tumor dataset. The results are evaluated based on the training time and the percentage classifier accuracy. The tables and the figures below present a detail and pictorial comparison of the filter based and wrapper based approach for the selecting of optimal features and classification of the selected features.

A. RBF KERNEL

APPROACH	ALGORITHM	CLASSIFICATION ACCURACY	BUILDING TIME
FILTER	CFS+SVM	64.52 %	0.05 sec
WRAPPER	NAÏVE BAYESIAN + SVM	64.52 %	0.746 sec

Table 4.1: The RBF Kernel

For the RBF kernel, the filter and wrapper approach presents a common result based on the classification accuracy but varies on the timing that was taken to build the model respectively. Based on timing, the filter approach gives a more timing reliant model than the wrapper approach.

B. POLY KERNEL

APPROACH	ALGORITHM	CLASSIFICATION ACCURACY	BUILDING TIME
FILTER	CFS+SVM	87.10 %	0.02sec
WRAPPER	NAÏVE BAYESIAN +SVM	89.03 %	0.028 sec

Table 4.2: The Poly Kernel

For the Poly kernel, the filter and wrapper approach present different results based on the classification accuracy and the timing that was taken to build the model respectively. Based on timing, the filter approach gives a more timing reliant model than the wrapper approach. However, based on the classifier accuracy, the wrapper approach, gives a more suitable model than the filter approach.

C. NORMALIZED POLY KERNEL

APPROACH	ALGORITHM	CLASSIFICATION ACCURACY	BUILDING TIME
FILTER	CFS+SVM	90.32 %	0.02 sec
WRAPPER	NAÏVE BAYESIAN+ SVM	85.81 %	0.532 sec

Table 4.3: The Normalized Poly Kernel

For the Normalized Poly kernel, the filter and wrapper approach present different results, based on the classification accuracy and the time that was taken to build the model respectively. Based on timing, the filter approach gives a more timing reliant model than the wrapper approach and based on the classifier accuracy, the filter approach gives a more suitable model than the wrapper which is often dataset dependent.

V. CONCLUSION

In this research work, a systematic feature selection reduction framework was investigated and a dual selection technique was used for selecting using microarray colon cancer datasets using the filter and wrapper approaches in order to solve the problem of high dimensionality reduction. The main purpose of the dual selection technique is to obtain an optimum reduced subset to a classifier algorithm so as to determine highest accuracy when classifying the dataset using a small subset of informative genes. The selected results was passed to the support vector machine which was subjected to three different kernels, namely, the Radial Bias Function (RBF) kernel, the Poly kernel and the Normalized Poly kernel. The classification accuracy of colon cancer data set gives a high classification result for the both selection approaches using the Poly kernel and normalized poly kernel. The same level of classification accuracy was obtained for the RBF kernel. The Normalized Poly kernel has the filter approach with the better classifier accuracy, because the scale of lung data set is larger, while the Poly kernel presented the wrapper approach with the highest classifier accuracy than the filter approach. Generally from the experimental point of view it was observed that the filter approach is more time computational intensive than the wrapper approach as the results obtained justify the result for this.

REFERENCES

[1] R. E. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, USA, 1957.
[2] S. Y. Kung and M. W. Mark (2009). Machine Learning in Bioinformatics, Chapter 1: Feature Selection for Genomic and Proteomic Data Mining, John Wiley & Sons, Hoboken, NJ, USA, 2009.
[3] J. Han (2005). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2005.
[4] D. M. Strong, Y. W. Lee, and R. Y. Wang (1997). "Data quality in context," Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.
[5] J.R. Quinlan (1986). "Induction of decision trees," Machine Learning, No. 1, 1986, pp.81-106.
[6] Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., and Staudt, L.M. (2000). Nature, pp.403-503.
[7] A. Bhattacharjee, W.G Richards, J. Staunton, C. Li.(2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adeno-carcinoma subclasses. Proc Natl. Acad. Sci. U S A. 2001; no. 98, pp. 13790–13795.
[8] L. Yu and H. Liu (2004), Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learning Res. 5 (2004), pp. 1205–1224.
[9] C. Ding and H. Peng (2003), Minimum redundancy feature selection from microarray gene expression data, in: IEEE Computer Society Bioinformatics, 2003.
[10] E. Xing, M. Jordan and R. Karp (2001). Feature selection for high-dimensional genomic microarray data, in: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2001.