New Feature Selection Method for High Dimensional Gene Data

M.N.F. Fajila and R.D. Nawarathna*
Department of Statistics and Computer Science
University of Peradeniya
Peradeniya 20400, Sri Lanka
*ruwand@pdn.ac.lk

***Abstract*— Dimensionality reduction (i.e., feature selection) is an essential technique in data science when handling high dimensional data such as cancer microarray samples. Cancer microarray experiments normally provide a large number of data which is assumed to contain many features, called, genes. However, genes can be either redundant or irrelevant, and thus be removed without incurring much loss of information. A small number of samples with a large number of genes is the major problem in microarray data analysis. In this study, a new machine learning method, namely, hybrid wrapper – filter feature selection is proposed for gene selection. This approach combines the genes selected by both filter and wrapper feature selection methods. Further, it uses a least priority feature elimination procedure where the genes with the lowest priority are eliminated. The propsoed technique is validated and evaluated on two microarray data sets namely, leukemia and colon cancer data sets. With gene selection performed by the proposed method, it helps to classify the leukemia microarray samples with perfect classification (100%) and to classify the colon cancer data set only with two misclassifications giving an accuracy of 90.5%. Results show that the proposed technique is extremely efficient in terms of the computational time too.**

Keywords; Classification, Dimensionality reduction, Feature selection; Gene selection; Microarray experiment.

I. INTRODUCTION

Cancer classification is an important step in the treatment of cancer. In the context of cancer, gene expression profiling, which is a technique used to query the expression of genes, has been used to classify tumors. One of the techniques is Deoxyribo Nucleic Acid (DNA) microarray technology, which provides thousands of genes simultaneously. A small number of samples with a large number of genes is the major problem in microarray data analysis. This may lead to a decrease in prediction accuracy and an increase in overfitting problems. To select relevant genes involved in cancer remains a challenge. Traditionally manual management of microarray data is impractical. Therefore, machine learning techniques are used to discover informative knowledge from these data. Gene selection is a dimensionality reduction algorithm (i.e., feature selection algorithm) [1], in which irrelevant and redundant genes are eliminated. Identifying a smallest and the most informative subset of genes is the goal of gene selection. Therefore, gene selection is widely used in microarray data analysis to reduce the large dimension and select the relevant genes involved in cancer to classify the cancers. Thus, gene selection improves the prediction accuracy in cancer classification. Several machine learning approaches have been already carried out for cancer classification in past decades. Voting machines and self-organizing maps (SOM) [2], support vector machines (SVMs) [3] and support vector machine recursive feature elimination (SVM-RFE) [4] are some of the approaches carried out so far. Although there are so many approaches have been developed, selecting the best subset of genes with a higher accuracy efficiently is still a challenging task.

II. MATERIALS AND METHODS

In this study, a new hybrid wrapper–filter feature selection method is proposed as summarized in Fig. 1. In the proposed approach, the genes selected using both wrapper and filter feature selections are combined as seen in Fig. 1. Gene selection is further refined using a least priority feature elimination procedure. Filtering can be used to reduce dimensionality and overcome overfitting. In the filter approach, a filter [5] along with a ranker search method is used to prioritize the genes. The major problem in the filter approach is the computation of a threshold by which features may be discarded from the ranking. One heuristic approach is known as the n-1 rule in microarray cancer analysis where n denotes the number of instances chooses the top n-1 genes to start the analysis [1].

The analysis is started according to n-1 rule and genes are removed continuously, until the highest performance for training sample is obtained using a naïve Bayes classifier [5]. Two types of filters, namely information gain [5] and

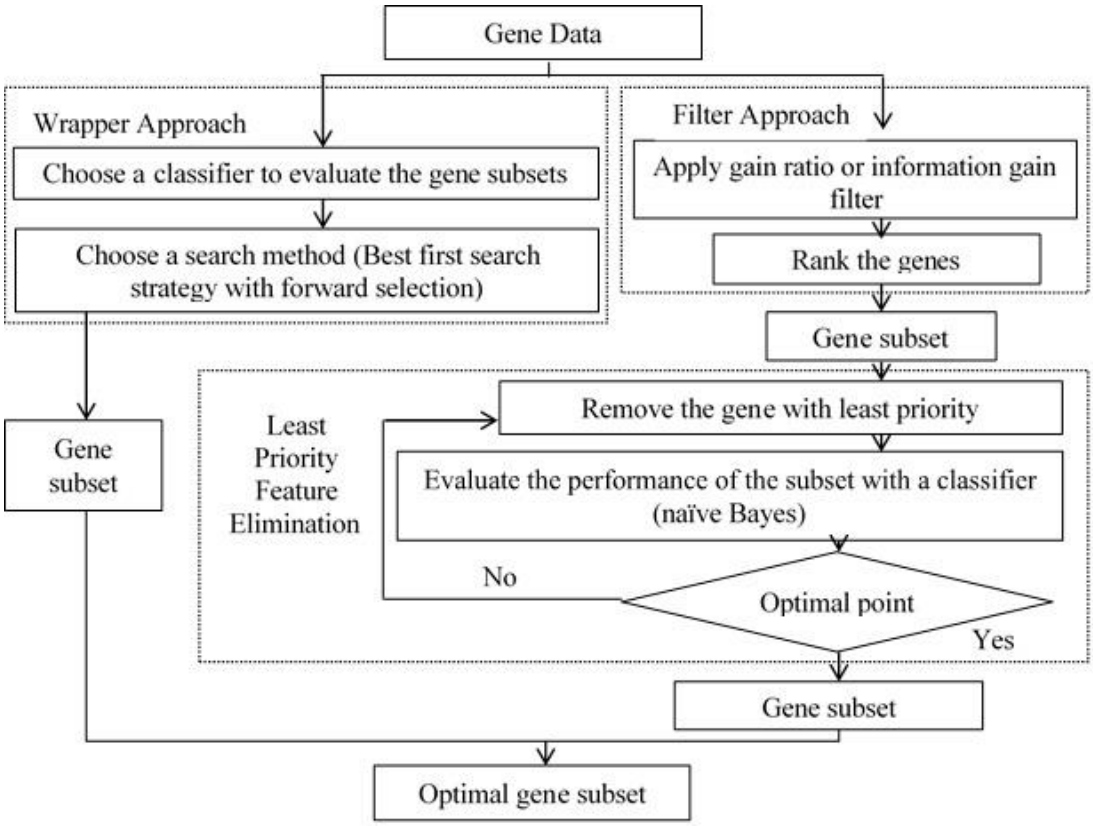

Figure 1. Steps in the proposed Hybrid Wrapper-Filter gene selection method

gain ratio [5] are used here. Information gain filter evaluates the worth of a feature by measuring the information gain with respect to the class. Gain ratio evaluates the worth of an attribute by measuring the gain ratio with respect to the class. A feature can be more informative in the presence of another feature. Thus, wrapper method selects all the possible combinations of feature subsets and finally provides the most informative subset which classifies the gene data with a higher accuracy. That is, the best gene subset, among all possible gene subsets, is evaluated according to a given classifier and a search strategy. The classifier subset evaluator [5] and the best first searching strategy along with forward selection are used for wrapper method. A classifier is used to estimate the merit of a set of features. "Merit" [6], in the case of the classifier subset evaluator, is either the number of incorrectly classified instances (ICC) for nominal classes or the root mean squared error (RMSE) (for numeric classes).

The least priority feature elimination is a process where the feature with the lowest priority is eliminated from the selected subset to evaluate the performance using naïve Bayes classifier. This step ensures the removal of the features with least priority to reduce the size of the feature subset while improving the classification accuracy. At the end of the evaluation, the optimal performance is checked. If the subset provides optimal performance than the previous subset, then that subset is selected. Or else, the next feature with the lowest priority is eliminated to check for the optimal point. Likewise, the best feature subset for each of the filter is selected after the optimal point is reached. The selected set of genes can be used for the cancer classification using a classifier such as naïve Bayes, support vector machine (SVM), decision tree [5] and so on. Fig. 1 shows all the steps in the hybrid wrapper–filter gene selection approach.

III. RESULTS AND DISCUSSION

Two data sets namely, colon cancer data set [7] consists of 62 instances each with 2000 genes and Leukemia data set [2] consists of 72 instances each with 7129 genes were used for the evaluation of the proposed method as shown in Table I.

TABLE I. GENE DATA SETS USED FOR THE EVALUATION

Data Set	**Number of Genes**	**Number of Instances**
Colon Cancer	2000	62
Leukemia	7129	72

Three classifiers, namely the naïve Bayes, the J48 decision tree learner and Sequential Minimal Optimization (SMO) support vector machine [5] were used as classifiers and WEKA machine learning software [8] was used for the implementation of the classifiers. For both data sets, first, the most informative genes are selected using proposed gene selection algorithm. Then, a cancer classification is performed with and without gene selection. Accuracy, Receiver Operating Characteristic (ROC) [5] value, the number of Correctly Classified Instances (CCI), the number of Incorrectly Classified Instances (ICI) values are reported and compared for each classifier.

In the analysis of colon cancer data set, 16 genes for gain ratio filter and 4 genes for information gain filter were selected. All the genes selected by information gain filter overlapped with that of gain ratio. The gene accession numbers of overlapping genes are M63391, R87126, M76378 and M26383 [7]. The gene M63391 overlapped with that of wrapper approach as well. Eight genes were selected in the wrapper approach. 0.032 merit value was obtained for wrapper approach when naïve Bayes classifier was used for feature selection. Altogether 23 genes were selected for colon cancer data set.

In the analysis of leukemia dataset, 30 genes for gain ratio filter and 8 genes for information gain filter were selected. All the genes selected by information gain ratio overlapped with that of gain ratio. Among these genes, Zyxin [7] overlapped with that of the wrapper approach as well. Two genes were selected in the wrapper approach. Zero merit value was obtained for wrapper approach when naïve Bayes classifier was used. Overall, 31 genes were selected for the leukemia data set.

The performances of the colon cancer and leukemia cancer classifications with and without gene selection are given in Table II and Table III, respectively. According to Table II and III, the results obtained indicate that the proposed approach has given a higher performance for classification when the gene selection is performed with the proposed approach. 90.5% accuracy with 97.1% ROC for colon data set and 100% accuracy with 100% ROC for leukemia dataset were obtained. It can be seen that the accuracy of the colon cancer classification has been increased up to 90% for the Naïve Bayes classifier. It is clear that the redundant genes in data sets may cause overfitting and a drop in the accuracy (52.4%) which is the case with Naïve Bayes classifier. For the colon cancer data set, the reported RMSE value for naïve Bayes classifier was 0.309 whereas for the leukemia data set the reported RMSE value was 0.0004 for naïve Bayes classifier. Both data sets have nominal classes and so merit value [6] denotes the error rate. Also, the classification time taken for colon data set and leukemia data set are 25.83 seconds and 136.66 seconds, respectively.

Moreover, the overlapped genes indicate that the selected genes are more informative for classification. Some of these genes have already been selected in other related studies [1], [2], [4]. For instance, Zyxin in leukemia has been given a higher priority in the gene subset selected by [1], [2] and [4] and as well by the proposed method.

IV. CONCLUSION

The proposed hybrid wrapper-filter feature (i.e., gene) selection method that uses wrapper and filter feature selection methods provides comparatively better results for gene selection. The overall idea of the method is to combine the most informative genes selected from both wrapper and filter feature selection methods. The method also uses a least priority feature elimination procedure to further enhance the informative subset of genes. In the filter approach, a filter along with a ranker search method is used to prioritize the genes. Two types of filters, namely, information gain and gain ratio are used. In the wrapper approach, the best gene subset, among all possible gene subsets is evaluated according to a given classifier and a search strategy. Wrapper method takes all possible combinations of feature subsets and finally selects the most

TABLE II. PERFORMANCE WITH AND WITHOUT GENE SELECTION FOR COLON CANCER DATA SET

Classifier	**Without gene selection**			**With gene selection**		
	Accuracy (%)	***CCI(ICI)***	***ROC (%)***	***Accuracy (%)***	***CCI(ICI)***	***ROC (%)***
Naïve Bayes	52.4	11(21)	56.7	**90.5**	**19(21)**	**97.1**
SMO	81.0	17(21)	88.2	90.5	19(21)	94.1
J48	76.2	16(21)	85.3	81.0	17(21)	78.7

TABLE III. PERFORMANCE WITH AND WITHOUT GENE SELECTION FOR LEUKEMIA DATA SET

Classifier	**Without gene selection**			**With gene selection**		
	Accuracy (%)	***CCI(ICI)***	***ROC (%)***	***Accuracy (%)***	***CCI(ICI)***	***ROC (%)***
Naïve Bayes	88.2	30(34)	86.8	**100.0**	**34(34)**	**100.0**
SMO	85.3	29(34)	82.1	85.3	29(34)	82.1
J48	91.2	31(34)	91.4	91.2	31(34)	91.4

informative subset as a feature can be more informative in the presence of another feature. Further, the proposed method uses a least priority feature elimination procedure where the genes with the lowest priority are eliminated.

An evaluation study conducted with leukemia and colon cancer microarray data shows that the proposed approach guarantees a perfect classification (100%) of the leukemia microarray samples. For colon cancer data set, the proposed method helps to classify only with two misclassifications giving an accuracy of 90.5%. Thus, the new method achieves good results with gene selection and guarantees reliable classification for new unclassified samples in cancer classification. Further, the results show that proposed technique is extremely efficient in terms of the computational time too.

REFERENCES

[1] Y. Wang, I. V. Tetko, M. A. Hall and E. Frank, "Gene selection from microarray data for cancer classification: A machine learning approach", *Comput Biol Chem*, vol. 29, pp. 37-46, Feb. 2005.

[2] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.H.H.C. Mesirov, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". *Science*, 286 (5439), pp. 531–537, Oct. 1999.

[3] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data". *Bioinformatics*, vol. 16, pp. 906–914, Oct. 2000.

[4] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Gene selection for cancer classification using support vector machines". *Machine Learning*, 46(1-3), pp. 389–422, Oct. 2002.

[5] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[6] "[Wekalist] what is the meaning of "Merit of best subset found" in the result of selecte attributes", [Online] Available: https://list.waikato.ac.nz/pipermail/wekalist/2008-November/041415.html [Accessed Sept. 2015].

[7] "Bioinformatics Research Group (Dataset Repository in ARFF (WEKA))", [Online] Available: http://eps.upo.es/bigs/datasets.html [Accessed: Aug. 2015].

[8] "Weka 3: Data Mining Software in Java", [Online] Available: http://www.cs.waikato.ac.nz/ml/weka/downloading.html [Accessed July 2015].