Cancer Classification using Ensemble Feature Selection and Random Forest Classifier

Nimrita Koul, Sunilkumar S Manvi
REVA University, Bangalore, Karnataka, India - 560064
Emailnk1@gmail.com

Abstract. High volumes of genomic data made available by high throughput gene expression sequencing technologies like next generation sequencing, microarray gene expression data have made it possible to develop models to computationally analyse this data and infer meaningful insights like presence of a disease, nature of disease, place of localization of the tumour in cancers etc. Since gene expression data is very high dimensional, each gene stands for one dimension, and has very small number of observations, it is imperative to apply feature selection on the data before using it for classification task. In this paper, we have proposed a method for classification of human cancer types by analysis of microarray gene expression data. We have used an ensemble feature selection algorithm for selecting subsets of 5, 10, 20 and 30 genes and applied random forest classifiers to obtain the classification accuracy and other performance parameters for comparison with existing solutions. We have been able to obtain 100% classification accuracy with just 5 genes on colon cancer data set with our algorithm.

Keywords – Feature Selection, Gene Expression Data, Machine Learning, Kernel PCA, Ensemble Techniques

1. INTRODUCTION

With the advancement in medical diagnostic technologies like DNA microarray technology [1], next generation sequencing etc. enormous volumes of genomic data has been generated. DNA microarray experiment produces the gene expression values of all thousands of genes in the genome of an organism being sequenced. So this data has very high number of dimensions and at the same time the number of observations or samples is very less.

This data can be analysed using computational analysis techniques like machine learning to understand various patterns in data. A study of such patterns can reveal an abnormality or a disease condition in the cell. This analysis can also reveal functional, regulatory relationships between different genes within a cell. It can also be used to study relationships between physiological functions and cellular interactions in an organism at the level of genes. After a microarray experiment, the raw data that is produced is pre-processed to perform background corrections, normalization, standardization, summarization [2, 3] etc. Once the preprocessing is over, the gene expression data sets are ready of analysis using machine learning techniques. The prominent tasks of machine learning that have been carried out by researchers on gene expression data sets are classification and clustering [4, 5, 6]. The tasks of classification and clustering are used to profile the gene expression data. Clustering aims at partitioning of genes into groups based on certain similarity measures in such a way that the genes which are related to each other by function or structure or effect are placed into same group and unrelated genes are placed in different groups. Classification is another important machine learning tasks that aims at labelling the activity of genes as normal or abnormal which can be used to diagnose a cell as normal or diseased. This can be used to classify patients according to health condition. Since gene expression data is a typical example of the data with very large number of
dimensions and very small number of samples, the curse of dimensionality [7, 8] often figures in computational processing of gene expression data, in addition to the problem of ‘large p and very small n’, there is an acute class imbalance in such data besides the noise and missing values. From clinical research, we know that expression of very few genes is responsible for onset of diseases like cancer [9], therefore the task of identifying these few genes from expression of entire genome is an important task. These few genes, once identified, have the ability of correctly classifying patients as normal or diseased without the computational processing of entire genome. In this paper, the focus is to classify gene expression profiles as cancerous or non-cancerous. We have worked with colon cancer dataset [14]. Objective of a classification task is labelling of a gene in the given gene expression dataset as being relevant to diagnosis or the disease or not. Once we have found the genes not relevant to diagnosis, we drop them and retain the genes with most influence on diagnosis. This task of selecting the genes which are highly related to diagnosis is known as feature selection [10]. Depending upon the nature of machine learning task – supervised, unsupervised or semi-supervised, there are various methods for feature selection. In supervised feature selection, we have the information about diagnosis or labels of each sample, therefore the task of feature selection algorithm is to identify those genes whose expression values are highly correlated with value of label. Once the features have been selected, the classification task can begin.

Since the objective of classification task is to identify correct labels for the samples in gene expression dataset, the computational algorithm has to analyze each expression profile for its relevance to labels present during training of the classifier. This analysis on thousands of genes is compute intensive and the curse of dimensionality can influence the results. Therefore, the feature selection task precedes the classification task while processing gene expression datasets.

1.1. Background

The process of feature selection is a subset selection problem, i.e. the processing carried out to select a minimum number of most relevant and mutually unrelated genes which are sufficient to correctly classify a sample from gene expression dataset. The important task of differentiating normal human cell from a cancerous cell can be carried out by computationally observing difference in values of expression levels of genes in a sample taken from a person. It is much faster and efficient to observe just the relevant genes than it is to observe all twenty four thousand genes present in the genome. Thus it is important to find out the subset of genes that perform best with some classifier algorithm. The measure of being best if often classification accuracy, true positive rate, precision, recall, sensitivity and specificity

The process of feature selection on original data set with Y number of features is the process of selecting a minimal subset with X number of features, such that X ⊆ Y and X minimizes the error in classification. [10,11,12] i.e., minimum number of samples are wrongly classified when basing the decision only on X. Error can be computed using a standard error function like mean squared error. Thus the process of feature selection is also a constrained optimization problem where we are trying to minimize classification error, in other words, we are trying to maximize classification accuracy. There can be constraints on the error function e.g., maximum permissible unit or error.

1.2 Advantages of Feature Selection

The advantages of feature selection are –

1. Removing the irrelevant features from the input dataset can lead to improvement in classification accuracy of the classifier.
2. Having only the relevant genes in the dataset can improve validation scores on test data.
3. This also increases generalizability of the classifier trained on this data.
4. Due to smaller dimensionality, the chance for overfitting is reduced.
5. Lower dimensionality of relevant feature set also reveals interesting aspects of the problem.
6. Lower dimensionality of input dataset, reduces the computation resources, time and memory required for processing.
7. Often, the classification task involve learning or search into the solution space, with less number of features, the search space is reduced.

Literature classifies feature selection methods as univariate or multivariate depending on whether each feature of the original dataset is considered individually to compute its relevance to class labels or a combination of features is considered at a time. While univariate feature selection considers genes individually, multivariate gene selection considers groups of genes at a time and therefore considers interdependencies and interrelations among genes.

Another way of classifying feature selection methods is as filter methods, wrapper methods, and embedded methods. This categorization is done based on the mechanism of validation used to identify relevance of a feature. Filter approaches exploit inherent properties of data for subset selection, these methods are independent of the predictor variables. The feature selection methods under the wrapper approach work on the principle of optimizing a predictor variables. The features that optimize value of predictor are retained in selected subset. The class of feature selection methods which fall under the category of embedded approach use a machine learning algorithm known as classifier as a part of the feature selection process. Each feature is evaluated by measuring its influence on the classification accuracy. The optimal subset of genes is one that maximizes the classification accuracy and at the same time the size of this subset should be minimum. Both wrapper and embedded methods have a drawback of overfitting and have high computational demands.

Feature selection methods which constitute filter methods use the information theoretic and statistical properties of data entropy, information gain, mutual information \([9]\), correlation coefficient, consistency, fast correlation to identify the best genes. Correlation coefficient and fast correlation, consistency are multivariate filter methods as they act on multiple genes at a time. Both correlation and fast correlation methods compute the correlation coefficient of a gene or feature with the class label, the genes with highest values of correlation are preferred. However, we also want to select the genes which in addition to having highest correlation with label have least correlation with other genes in the feature set. We wish to retain most mutually unrelated genes that are most related to labels. A gene is considered highly relevant if it is highly correlated to the class of the sample and correctly predicts it and at the same time is uncorrelated to other genes in the input set of features.

Among important univariate methods of feature selection are mutual information \([9]\) and information gain. In this method, the gain in information content is computed for each gene with respect to the label. Information gain is a measure of decrease of entropy due to inclusion of a feature. It is an important metric in training of decision tree based feature selectors and classifiers. For each feature, information gain can be computed by observing the difference between entropy of a dataset with or without the feature that is being evaluated. Mutual information is a related metric which is a measure of statistical inter-dependence of two variables. Mutual information indicates the amount of information that can be inferred about a feature from another feature. More the mutual information between two features, more is the redundancy.

Other prominent filter approach to feature selection is known as minimum redundancy maximum relevance. In this method, the genes which are most relevant to the class label and at the same time most dissimilar with other genes are selected in the subset. Many methods of feature selection perform the analysis of variates for differentiation of classes through identification of significant statistical differences between means of classes using metrics like sum of squares,
Fisher score, Laplacian score etc. Through calculation of these metrics the feature subsets are identified which maximize inter class dissimilarity measure and minimize within class dissimilarity measure and are able to maintain their locality.

Rest of this paper is organized as follows – Section II covers the survey of literature and the methods proposed for feature selection from gene expression data, Section III presents the proposed algorithm, in Section IV covers the results and discussion, Section V concludes the paper.

2. RELATED WORK

In this section, we cover the study of feature selection methods proposed in literature. The methods we present here are prominently applied to high dimensional gene expression data. These methods are primarily filter methods which are suitable for handling gene expression datasets.

In [1] authors demonstrated for the first time that accurate classification could be done by using computational approaches for analysis of gene expression datasets. The authors worked on Leukaemia data set and successfully classified it into two sub classes of leukaemia with very few relevant genes. The work in [2] presents the use of support vector machine as feature selection tool for the task of cancer classification from gene expression datasets. In [3], the researchers have employed regression technique called as least squares regression for identification of most relevant genes from the biological data for various classification tasks. Authors in [4], used a method called minimum redundancy maximum relevance as a feature selection approach and artificial neural networks as classifiers for classification tasks on biological data. Authors in [5], have employed evolutionary approach and fuzzy logic based method for feature selection. They have used Bee colony optimization combined with fuzzy logic for feature selection from gene expression data. Authors in [6], applied distributed correlation methods for feature selection from gene expression datasets for the classification problem. Paper [7] is a survey of prominent feature selection approaches for the task of classification of cancers using gene expression as input data. In [8], researchers have employed logistic regression as feature selection technique for selection of relevant genes for gene expression datasets. Authors in [9], applied a pipeline of mutual information and genetic algorithm to produce an ensemble algorithm for feature selection for cancer classification using gene expression profiles. In [10], the authors have used kernel SVM as feature selection method on cancer gene expression data sets. In [11], authors used a hybrid algorithm consisting of an ensemble of cellular automata and an evolutionary algorithm called ant colony optimization for obtaining the optimal feature subset from microarray gene expression data. Authors in [12] compute best ‘r’ genes by dividing dataset into subsets. From each subset top genes are selected and finally all combined together to get the best overall genes of the dataset. The process is repeated till an optimal subset is obtained. Authors in [13], have applied an ensemble of Support vector machine and recursive feature elimination as a method for gene selection from gene expression datasets. In this study, we have experimented on the Colon cancer dataset [14], containing 60 samples, 2000 genes with two classes – normal and tumour.

3. PROPOSED METHOD

The proposed algorithm for classification is presented in this section. This method is an ensemble of mutual information and kernel PCA. The algorithm is as follows –

Algorithm 1: MIPCA Gene Selection

| Algorithm 1: MIPCA Gene Selection |
|-----------------------------------|
| **Input:** I - Set of Gene Expression Data |
| **Output:** O - Subset of most relevant Genes |
| 1. Compute the mutual information of all genes in I |
| 2. Compute score of I with RFC with depth value of 4 |
3. Select top ranked 5, 10, 20 and 30 gene sets with highest MI values as X5, X10, X20, X30
4. Apply Kernel PCA to I and X5, X10, X20, X30 generating new Xs.
5. Apply RFC-4 to new Xs.
6. Compute the performance parameters.
7. Return the subset with maximum value of score

Algorithm 1: Gene selection Approach

Experimental set up consisted of a Windows 10 machine with Intel i7, 12 cores processor. Programming was done in Python 3. The dataset used is the Colon cancer dataset [14]. The dataset contains 60 samples, 2000 genes with two classes – normal and tumour.

We executed the experiment 20 times with each with full dataset, then with reduced data sets of lower number of features. We selected best 5, best 10, best 20 and best 30 genes using the algorithm presented above. The average values for all parameters were used for graphing and drawing conclusion. The classification was performed using Random Forest classifier with a depth of 4. During the feature selection phase, we used mutual information as first level of ranking of genes and kernel PCA as the second level of feature selection.

Fig.1. Schematic representation of the proposed scheme
Figure 1 shows the diagrammatic representation of the proposed algorithm. Each level shows the flow of data from top to performance parameters – classification accuracy and training time obtained at bottom end.

4. Results and Discussion

In the experiment, we found the pairwise mutual information in the features in full dataset [14]. The classification accuracy was noted without this computation. The top ranked 5, 10, 20 and 30 genes were noted and classification was performed using just the reduced feature sets, and corresponding performance noted. Then kernel PCA was applied on the full input set and the reduced sets and again the accuracy of classifier noted for comparison. The input data is standardized and normalized. Classification accuracy with the proposed method is 100% with a subset of 20 genes.

| Number of Features | MIRFC | PCARFC | MIPCA-RFC | RFC |
|--------------------|-------|--------|-----------|-----|
| Full (2000 genes)  | 100   | 98     | 100       | 100 |
| 5                  | 93    | 91     | 98        | 96  |
| 10                 | 92    | 95     | 98        | 98  |
| 20                 | 96    | 98     | 100       | 99  |
| 30                 | 97    | 96     | 100       | 99  |

Table 1. Comparison of Classification Accuracy

Table 1 shows that we get 100 % classification accuracy with full feature set as well as with best 10 features. Therefore these 10 features can be considered relevant to colon cancer.

Figure 2 shows the comparison of classification accuracy obtained with original data set with 2000 features under four experimental conditions – first, using mutual information as feature selection algorithm and random forest as classifier, second, PCA as feature selector, RFC as classifier, third, proposed algorithm ensemble of MI and PCA as feature selector and RFC as classifier, fourth, using
RFC without explicit feature selection. The classification accuracy obtained with 30 selected genes is better than with 5, 10 or 20.

Table 1 shows the values of classification accuracy in percentages for the four experimental conditions mentioned above.

![Figure 3. Comparison of Training Times required with different number of feature](image)

Figure 3 shows the training times required to fit the classifier with different sizes of feature set. It can be observed that the training times are highest for full feature set consisting of 2000 genes present in original data set.

5. CONCLUSION

Filter methods of feature selection prove to be most suitable for the feature selection from high dimensional gene expression data. These also are less compute intensive as compared to wrapper and embedded methods for this task. Among the classification algorithms, SVM and RFCs have been demonstrated to work very well with gene expression data. In this paper, we presented an ensemble technique for feature selection and RFC was used for classification of the data with and without feature selection. The algorithm identified 10 genes which could accurately classify the input dataset. Thus the minimal subset of genes has size 10 for accuracy of 100%. As a future work, the proposed algorithm has to be applied on other types of data and performance observed.

ACKNOWLEDGEMENT

Authors would like to thank Department of Science and Technology (DST), Government of India, for financially supporting this work under the scheme DST-ICPS 2019.

REFERENCES

[1] G T.R. Golub et. al., 1999, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, pp. 531–537.

[2] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, 2002, Gene selection for cancer classification using support vector machines, *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp 1-14

[3] Lê Cao, K., Boitard, S. & Besse, P, 2011, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC*
Bioinformatics, vol. 12, no. 253, https://doi.org/10.1186/1471-2105-12-253

[4] Masud Rana M., Ahmed K., 2020, Feature Selection and Biomedical Signal Classification Using Minimum Redundancy Maximum Relevance and Artificial Neural Network, Proceedings of International Joint Conference on Computational Intelligence Algorithms for Intelligent Systems, Springer.

[5] Nancy S.G., Saranya K., Rajasekar S., 2020, Neuro-Fuzzy Ant Bee Colony Based Feature Selection for Cancer Classification, Springer Innovations in Communication and Computing. Springer, Cham 2020.

[6] Alok Kumar Shukla, Diwakar Tripathi, 2020, Detecting biomarkers from microarray data using distributed correlation based gene selection, Genes & Genomics. https://doi.org/10.1007/s13258-020-00916-w

[7] Nada Almugren, Hala Alshamlana, 2019, Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification, IEEE Access, vol. 7, pp. 75833-44 10.1109/ACCESS.2019.2922987

[8] Zakariya Yahya Algarmal, Muhammad Hisyam Lee, 2019, A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification, Advances in Data Analysis and Classification, vol. 13, pp: 753–771, https://doi.org/10.1007/s11634-018-0334-1

[9] M. Jansi Rani, D. Devaraj, 2019, Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification, Journal of Medical Systems, vol. 43, no, 235, https://doi.org/10.1007/s10916-019-1372-8

[10] Medjahed SA, Saadi TA, Benyettou A, Ouali M, 2017, Kernel-based learning and feature selection analysis for cancer diagnosis, Applied Soft Computing, vol. 51, pp 39–48

[11] Fatemeh Vafaee Sharbaf, Sara Mosafer, Mohammad Hossein Moattar, 2016, A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization, Genomics, vol. 107, pp.231–238, http://dx.doi.org/10.1016/j.ygeno.2016.05.001

[12] S. Miyano, S. Imoto, A. Sharma, 2012, A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754–764

[13] P. Mundra, J. Rajapakse, 2010, SVM-RFE withmRMR filter for gene selection, IEEE Transactions on Nano. Biosci, vol. 9, no. 1, pp 31-37

[14] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ, 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc Natl Acad Sci., vol. 96, pp. 6745-6750. A reference