Indian Journal of
Engineering

ISSN 2319 – 7757 EISSN 2319 – 7765

Gene Selection of Microarray Data Classification and Prediction for Colon-Rectum Cancer using Integer Programming

Shenbaga Ezhil S[1], Vijayalakshmi C[2]

1. Department of Mathematics, Sathyabama University, E-mail: shenbaga_ezhil@rediff.com
2. Department of Mathematics, Sathyabama University, E-mail: vijusesha2002@yahoo.co.in

Received 29 November; accepted 26 December; published online 01 January; printed 16 January 2013

ABSTRACT

Cancer is a family of diseases originated from genetic abnormalities which may be inherited. The classification of cancer is essential for administrating the most effective treatment, and has been traditionally based on the analysis of its morphological appearance. Extracting useful information from expression levels of thousands of genes generated with microarray technology needs a variety of analytical techniques. A new mixed integer programming model is formulated for this purpose using colon rectum cancer data base. Artificial Neural Network (ANN) is a branch of computational intelligence that employs a variety of optimization tool to "learn" from past experience and use that prior training to classify a new data, identify new patterns or prediction of such of colon rectum cancer.

Keywords: Mixed Integer Programming, colon rectum cancer, genetic microarray technology, classification, ANN.

1. INTRODUCTION

The objective of this paper is to use minimum number of genes (or) bio markers to classify tissue samples as accurately as possible into known groups. A novel mixed integer programming model is formulated to represent and to solve gene selection and tissue classification problem. However the high dimensionality and small sample size of microarray data also poses severe challenges to filter approaches in terms of effectiveness. Some of the recent researchers focused on these challenges. Mathematical programming approaches for classification analysis outperform parametric methods when the data depart from assumptions underlying these methods.

Mixed Integer Programming model for classification have been developed in the last two decade. These studies open new avenues for disease gene identification and biomarker's discovery, which can be used for diagnosis and for drug efficacy and toxicity assesments. Chillagayan et al., (2002) and Szabo et al., (2002) proposed the ways to select subsets of genes to use in the classification of tissue samples. These results show that the mathematical programming approach can rival or outperform traditional classification methods.

2. COLON RECTUM CANCER CASE STUDY

The colon cancer dataset was originally analyzed by Alon et al., (1999). This dataset contains expression levels of 2000 genes with highest minimal intensity across 40 tumor and 22 normal colon tissues. The data is available from R package "dprep". Pre-filtering according to t-test left 851 genes that were significantly differentially expressed (p-value $\leq$ 0.1) for further gene selection.

3. GENE SELECTION FOR MICROARRAY DATA SETS

In this we consider the data base of colon cancer study. The colon cancer dataset was originally analyzed by Alon et al. (1999). This dataset contains expression levels of 2000 genes with highest minimal intensity across 45 tumor and 20 normal colon tissues. Pre filtering according to't' test left 859 genes that were significantly differentially expressed (p value $\leq$ 0.1) for further gene selection.

Based on the degree of dependency selecting good features on evaluation measures and classification algorithm measures and classification algorithm, we divide a good subset into three parts:

(i) Features those are absolutely necessary for classification.

(ii) The features that can be chosen based on properties of data set. Features that can be found using a classification algorithm as evaluation measures. We name them as dependent features.

4. THE MIXED INTEGER PROGRAMMING MODEL

Let F = {1, 2, ..., f} denote the set of the indices of the different genes in G_t and g_i^- and g_i^+, i $\in$ F, the vectors of the expression levels of gene i for normal and tumor tissues, respectively. For each gene i $\in$ F, define two binary variables

$$z_i^- = \begin{cases} 0, & \text{if the profile } g_i^- \text{ properly identifies normal tissues} \\ 1, & \text{otherwise.} \end{cases} \qquad (1)$$

Shenbaga Ezhil et al.
Gene Selection of Microarray Data Classification and Prediction for Colon-Rectum Cancer using Integer Programming,
Indian journal of engineering, 2013, 2(3), 11-13,

$$z_i^+ = \begin{cases} 0, & \text{if the profile } g_i^+ \text{ properly identifies tumor tissues} \\ 1, & \text{otherwise.} \end{cases} \qquad (2)$$

Indicating if the expression profiles of i correctly characterize the state specified by the class value of the corresponding tissues. If the gene discriminate function takes the form w′g – b = 0, where w defines the orientation of the hyper plane is the s-dimensional space R^s and b its offset from the origin, the following mixed integer optimization problem can be formulated

$$\min \sum_{i=1}^{\infty} (z_i^- + z_i^+) \qquad (A)$$

Subject to the condition

$$w' g_i^+ - b \geq -Q\, z_i^+, \quad i \in \mathcal{F}, \qquad (3)$$

$$w' g_i^- - b < Q\, z_i^-, \quad i \in \mathcal{F}, \qquad (4)$$

z–, z+ are binaries, w, b free,

Where Q is a sufficiently large constant scalar, and constraints (3) and (4) set the values of the binary variables z_i^+ and z_i^-, i ∈ F.

From the solution of the problem (A), obtained by a truncated branch-and-bound procedure, it is possible to find a set of genes useful for discriminating between normal and tumor tissues. In particular, for each gene i ∈ F the following measure, termed classification score, is computed.

$$CS_i = \delta_i^- + \delta_i^+, \qquad (5)$$

where δ_i^- and δ_i^+ represent the Euclidean distances of patterns g_i^- and g_i^+ from the separating hyper plane.

5. ANN

ANN is a branch of computational intelligence that employs a variety of optimization tools to learn from past experiences and use this prior training to predict and identify new patterns. In this neural network models have been used for the prediction of Colon Rectum Cancer. ANN is a network that simulates our human brain functions. It is composed of parallel computing units called Neurons. These neurons can be connected in various ways to form different Neural Network architectures. The most popular architecture is the multi-layer perception (MLP). It consists of two or more layers of neuron in which the layers are connected in sequential manner. Each neuron in turn is connected to other neurons in the different layer by weighted path ways. Signals are sent through these pathways to the other neurons. Each neuron sums the weighted signals and transforms the resulting signal as the output of the neuron using an activation function. The output signal is then sent to the other neurons in the subsequent layers. The first layer of the network is called the input layer which receives signals from the data entering the network. The last layer is called the output layer which generates the outcome to the outside world.

6. METHODOLOGY

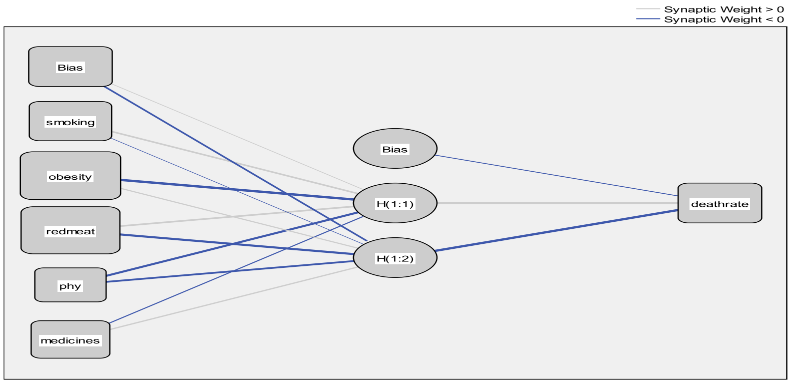
Artificial Neural Network (ANN) is a branch of computational intelligence that employs a variety of optimization tool to “learn” from past expenses and use that prior training to classify new data, identify new pattern and predict. In this study, a multi layer network with back-propagation (also known as multi layer perception) is used. Preprocessing the input data set for a knowledge discovery goal using data mining approach usually consumes the biggest portion of the effort devoted the entire work. In this work Neural network models have been used for the diagnosis and prediction of colon rectum cancer. Colon rectum cancer microscopic and clinical tests reports are collected. We have developed a set of tools to extract and clean up the raw SEER data to analyze the death rate.

A simple analysis shows that the SEER data has missing information in the field of Extent of Disease (EOD) and Site Specific Surgery (SSS) fields for almost half of the records. The SSS field usage has changed after 1998. Instead of the regular field, the information is split in five other fields. A mapping scheme from new SSS to old SSS is developed to fill the missing SSSS fields. The EOD field is composed of five fields including the EOD (Extent of Death) code. These fields are size of tumor, number of positive nodes, number of nodes and number of primaries unlike [6] we have included three fields

1. Survival Time Recode (STR)
2. Vital Status Record (VSR)
3. Cause of Death (COD)

The STR field ranges from 0 to 120 months in the SEER data base.

7. NETWORK DIAGRAM



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

8. RESULTS AND DISCUSSION

This table reports the running time for each feature selection algorithm. The number of genes was selected by each feature selection algorithm. We can see that half H FW on averages selects the smallest number of genes. Also it reports the leave-one-out accuracy by C 4.5.

Algorithm	HFW$_{C4.5}$	Half-HFW	FCBF
Running Tissue (s) for each feature selection Algorithm	59.83	0.91	1.14
Number of genes selected by each feature selection Algorithm	11	5	14
Validation accuracy of C4.5 on selected genes for each feature selection method	90.32	85.48	88.71
Leave one out cross validation accuracy selected genes for each feature selection method	75.81	90.32	77.42

9. CONCLUSION

Hence the classification of the data sets is consistently based on a very small number of genes, compared with the original number of features describing the sample tissues. Extensive experiments on microarray data have demonstrated the superior performance of our approach. The experimental results show that our model does not include records with missing data. Hence we would like to try survival time prediction of colon rectum cancer is seriously low than respiratory cancer.

REFERENCES

1. Wai-Ho Au, Keith CC, Chan Andrew KC, Wong, Yang Wang. Attribute Clustering for Grouping, Selection, and classification of gene expression data. *IEEE/ACM transactions on computational biology and bioinformatics*, 2005, 2(2)
2. David A. Bell, Hui Wang. A Formalism for Relevance and Its Application in Feature Subset Selection. *Machine Learning*, 41,175–195, 2000
3. Bobashev GV, Das S, Das A. Experimental design for gene microarray experiments and differential expression analysis. *Methods of Microarray Data Analysis II*, 2001
4. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *In Proceedings of the Computational Systems Bioinformatics Conference*, 2003, 523–529
5. Dougherty ER. Small sample issue for microarray-based classification. *Comparative and Functional Genomics*, 2001, 2, 28–34
6. Golub TR et al. Molecular classifications of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286(5439), 531–537
7. Hall M. Correlation-based feature selection for discrete and numeric class machine learning. *In Proceedings of the 17th International Conference on Machine Learning*, 2000, 359-366
8. Kohavi R, John G. Wrappers for feature subset selection. *Artif. Intell*. 1997, 1-2, 273–324
9. Ramaswamy S, Tamayo P, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 2001, 98(26), 15149–15154
10. Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 2003, 53, 23–69
11. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 1995, 270, 467–470
12. Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Slowinski R ed. Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992, 331–362
13. Usama M. Fayyad and Keki B. Irani, On the Handling of Continuous-Valued Attributes in Decision Tree Generation, Machine Learning, vol. 8, page 87–102, 1992.
14. Witten I, Frank E. Data Mining – Pracitcal Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann Publishers, 2000.
15. Wu Y, Zhang A. Feature selection for classifying high-dimensional numerical data. *In IEEE Conference on Computer Vision and Pattern Recognition*, 2004, 2, 251–258
16. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learning Res*. 2004, 5, 1205–1224