

# Gene Reduction for Cancer Classification using Cascaded Neural Network with Gene Masking

Raneel Kumar, Krishnil Chand, Sunil Pranit Lal

School of Computing, Information, and Mathematical Sciences  
University of the South Pacific

Suva, Fiji

{raneel.kumar, krishnil.chand, sunil.lal}@usp.ac.fj

**Abstract.** This paper presents an approach to cancer classification from gene expression profiling using cascaded neural network classifier. The method used aims to reduce the genes required to successfully classify the small round blue cell tumours of childhood (SRBCT) into four categories. The system designed to do this consists of a feedforward neural network and is trained with genetic algorithm. A concept of ‘gene masking’ is introduced to the system which significantly reduces the number of genes required for producing very high accuracy classification.

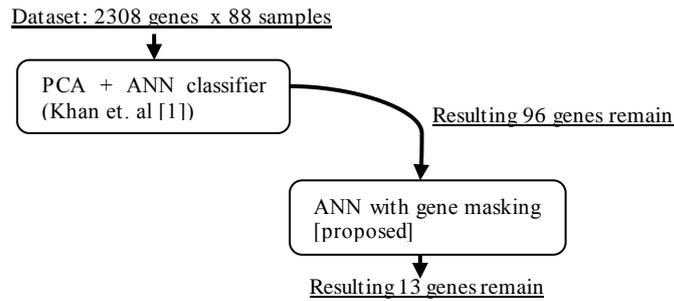
## 1 Introduction

Early cancer detection is important for the proper treatment of it. However some cancers cannot be easily identified and classified by traditional clinical means. Traditional clinical methods include diagnosis by X—Ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and ultrasonography [1]. Microarray gene profiling is a new way used to improve the accuracy of cancer classification. Microarrays can simultaneously measure the expression level of thousands of genes within a particular mRNA sample [2] but this is a difficult task due to the high dimensionality of gene expression data.

Dimensionality reduction can also be seen as the process of deriving a set of degrees of freedom which can be used to reproduce most of the variability of a data set [3]. Researchers have been involved in reducing the high dimensionality of the genes and at the same time preserving the features within the genes that would give a significant increase in accuracy of the classification process.

In this paper we focus on classification of the small round blue cell tumor (SRBCT) into four classes namely neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). Khan et. al [1] used principal component analysis (PCA) in training artificial neural networks (ANN) to progressively reduce the dimensionality of the SRBCT dataset from 2308 genes to 96 genes. Meanwhile Tibshirani et. al[4] applied nearest shrunken centroid classifier to the same dataset and showed reduction in the number of genes used to 43 with 100% accuracy. The nearest shrunken centroid classifier is essentially an extension of the nearest centroid classifier whereby features which are noisy and have little variation from the overall mean are eliminated using shrinkage factor.

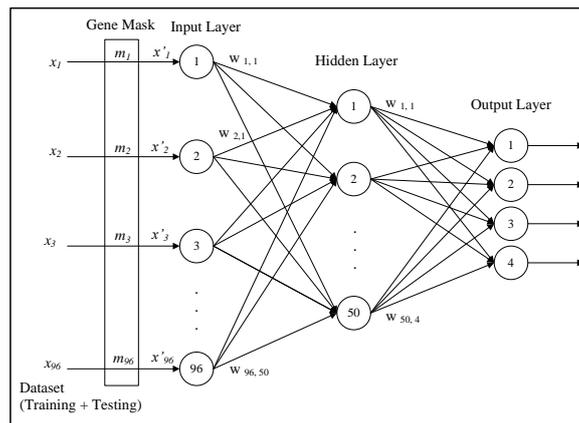
In this paper we extend the classifier proposed by Khan et. al [1] by incorporating a cascaded neural network classifier (Fig. 1) trained using genetic algorithm which leverages on our proposed concept of gene masking to further eliminate the number of genes required for accurate classification.



**Fig. 1.** Conceptual overview of the cascaded classifier.

## 2 Proposed Cascaded Classifier

The cascaded classifier (Fig. 2) has been implemented as a feedforward neural network [5] with 96 inputs expressing the relative red intensity of the corresponding genes obtained from the previous block (Fig. 1). The output layer for the neural network consists of four neurons corresponding to the four cancer types to be classified in this study. The outputs of the four neurons are compared, and final prediction is the cancer type corresponding to the neuron with the largest output. The number of neurons in the hidden layer [6] as well as the choice identity activation function [7] has been determined by empirical methods [8].



**Fig. 2.** Gene masking and classifier design.

## 2.1 Feature Selection by Gene Masking

The fundamental concept behind gene masking is that we should be able to eliminate genes which are not important for classification without adversely affecting the classification accuracy. However evaluating classification accuracy for all possible gene elimination combinations is computationally intensive for large number of genes. Therefore we use binary coded genetic algorithm to evolve optimal gene mask. The gene mask is a binary string with length equal to the number of genes being considered. Each gene data  $x_i$  is multiplied with the corresponding mask  $m_i$  to modify the network input  $x'_i$  as follows:

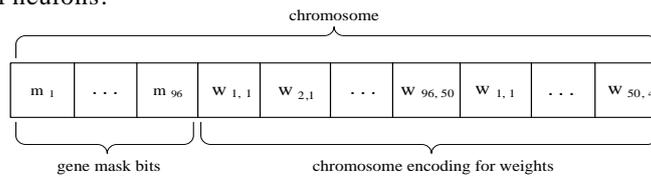
$$x'_i = \begin{cases} x_i, & \text{if } m_i = 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

## 3 Training the Classifier with Genetic Algorithm

The learning in neural network takes place by the adjustments of randomly initialized weights such that the classification error is minimized. We trained the neural network using genetic algorithm (GA) to search for optimal weight configuration which minimizes the classification error. GA [9] is a stochastic algorithm based on evolutionary ideas of natural selection and yields itself naturally to realizing the concept of gene masking.

### 3.1 Chromosome encoding

Chromosome encoding is the representation of the actual problem into a data structure which can be interpreted by GA. We used binary encoding to represent the gene mask and the weights for the neural network (Fig. 3). The length of the chromosome is 5096 bits where the first 96 bits are for gene masking, the next 4800 bits are weight values of the links from the neurons of the first layer to the neurons of the second layer while the next 200 bits are the weights value for the hidden layer neurons to the output layer neurons.



**Fig. 3.**Representation of a chromosome for the GA with  $w$  being weight encoding and  $m$  being genes mask

### 3.2 Fitness Function

The quality of solution represented by a chromosome is captured by its fitness value. Depending on the fitness function those chromosomes which pass the required criterion have more chance to enter the next step in the solution search process. The fitness function formulated for this classification problem takes into account the

accuracy of classification, that is, the success of classification of cancer into the right categories, and simultaneously maximizes the number of genes eliminated. The generalized fitness function for the system is given by:

$$\text{Fitness} = (\alpha * \text{Accuracy}) + (\beta * \text{Genes Eliminated}) , \quad (2)$$

where  $\alpha$  is the accuracy to gene elimination ratio in the range (0,1). In our study we set  $\beta = (1 - \alpha)$  to reduce the computational effort required to tune both the parameters.

## 4 Experimentation

### 4.1 SRBCT Dataset

The dataset [1] originally consisted of the expression of 6567 genes measured over 88 samples and was obtained through the cDNA microarray technology. After filtering out noise, 4259 genes were eliminated and the remaining 2308 genes were used as input to the classification scheme proposed by Khan et. al [1]. Our study uses the 96 genes output from [1] as the starting point. The 88 samples are divided in 63 training and 25 testing data samples. The training samples comprise of four tumor types: 8 Burkittlymphoma (BL) samples, 23 Ewing sarcoma (EWS) samples, 12 neuroblastoma (NB) samples and 20 rhabdomyosarcoma (RMS) samples. While the testing samples comprise of 6 BL samples, 3 EWS samples, 6 NB samples and 5 RMS samples with 5 non-SRBCT samples discarded in this study.

### 4.2 Training and Testing Phases

The training and testing procedure used in this study is captured in Table 1.

**Table 1.** Training and Testing Phases

Training Phase	Testing Phase
<ol style="list-style-type: none"> <li>1. A sample, <math>s</math> from 63 training sample is passed through the classifier with initial <math>\alpha = 0.01</math></li> <li>2. After an epoch, sample is classified into one of the four cancer types.</li> <li>3. Classification is compared with the actual result to get the accuracy</li> <li>4. This is done for all 63 samples to get fitness. The fitness considers the gene elimination ratio and classification accuracy</li> <li>5. The training is run for 20,000 epochs, using the fitness to guide GA in selecting optimal ANN weights for minimizing the classification error.</li> <li>6. Steps 1 to 5 is carried out for all <math>\alpha</math> in <math>A = \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}</math> such that we get 100% accuracy with maximum number of genes eliminated.</li> <li>7. For each <math>\alpha</math>, 10 runs are carried out by repeating steps 1-6, the best result is chosen for comparison</li> </ol>	<ol style="list-style-type: none"> <li>1. The best value of <math>\alpha</math> for testing is selected by choosing the smallest value of <math>\alpha</math> which produced 100% classification accuracy during training.</li> <li>2. For the chosen <math>\alpha</math>, we then obtain the best evolved chromosome during training.</li> <li>3. With the <math>\alpha</math> constant and the chromosome fixed we pass each sample from the 20 test samples through the classifier</li> <li>4. The predicted cancer type is compared with the actual cancer type to get the classification accuracy for all the 20 test samples.</li> </ol>

## 5 Results

The weight parameter  $\alpha$  assigns relative importance to the accuracy and number of genes eliminated. Large values of  $\alpha$  (close to 1) results in high accuracy at the expense of smaller number of genes eliminated (Fig. 4a). On the other hand  $\alpha$  values close to 0 eliminate larger number of genes but the accuracy is sacrificed (Fig. 4b). Therefore in selecting the best value of  $\alpha$  for testing, we choose the smallest value of  $\alpha$  which produced 100% classification accuracy during training. Accordingly we choose  $\alpha = 0.1$ . With  $\alpha$  set to 0.1, all the 20 test samples were correctly classified with 100% accuracy using only 13 genes.

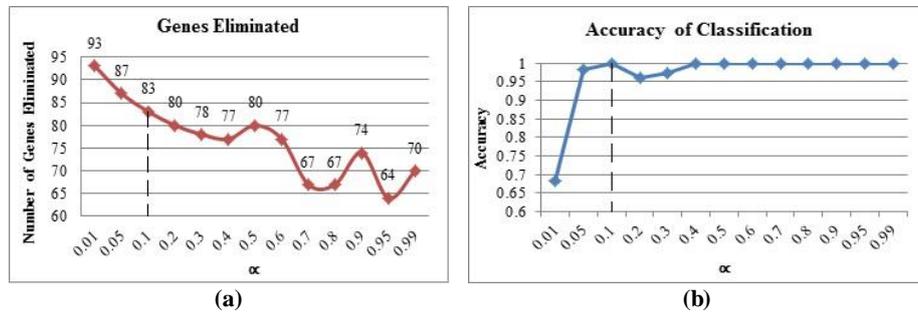


Fig. 4. Training results for various setting of  $\alpha$  parameter

### 5.1 Comparison of Results

The 13 genes which resulted in 100% classification of the test samples were noted and compared with [10]. Table 2 shows 54% of the genes in this experiment were also present in the comparator experiment.

**Table 2.** The 13 selected genes.

Gene	Image ID	Present in [10]
1	296448	✓
2	841641	✗
3	43733	✓
4	629896	✗
5	866702	✓
6	52076	✓
7	563673	✓
8	1469292	✗
9	1409509	✓
10	756556	✗
11	377671	✗
12	325182	✓
13	755599	✗

**Table 3.** Comparison of results obtained.

Method (Classifier)	Genes Remaining
PCA, MLP Neural Network [1]	96
Nearest Shrunken Centroid [4]	43
Gene Masking + ANN (this paper)	13

Comparison of results obtained on the SRBCT dataset with research methods reported in the literature show marked improvement in the number of genes required for accurate classification (Table 3).

## 6 Conclusion

In this paper we have extended the research reported in [1] by incorporating a cascaded neural network classifier trained using genetic algorithm. By applying gene masking, the learning algorithm was able to significantly reduce the number of genes required for accurate classification. The results show that the proposed system was able to achieve 100% classification accuracy on the SRBCT dataset using only 13 genes. Future work can be done to validate our approach on larger datasets (10K + features).

## References

1. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Perterson, C., and Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. 7, 673-679 (2001)
2. Sarhan, A.M.: Cancer Classification based on Microarray Gene Expression Data Using DCT and ANN. *Journal of Theoretical and Applied Information Technology*. 6, 208-216 (2009)
3. Ghodsi, A.: Dimensionality Reduction A Short Tutorial, Technical Report 2006-14., Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, pp. 5-6 (2006)
4. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*. 99, 6567-6572 (2002)
5. Rani, D.K.U.: Analysis of Heart Diseases Dataset Using Neural Network Approach. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*. 1, (2011)
6. Karsoliya, S.: Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. *International Journal of Engineering Trends and Technology*. 3, 714-717 (2012)
7. Karlik, B., Olgac, A.V.: Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence And Expert Systems*. 1, 111-122 (2011)
8. Singh, S., Chand, A., Lal, S. P.: Improving Spam Detection Using Neural Networks Trained by Memetic Algorithm. In: *Proceedings of 5<sup>th</sup> IEEE International Conference on Computational Intelligence, Mathematical Modeling and Simulation*, Seoul, Korea, pp. 55-60 (2013)
9. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. (1989)
10. Bair, E., Tibshirani, R.: Machine Learning Methods Applied to DNA Microarray Data Can Improve the Diagnosis of Cancer, *Special Interest Group in Knowledge Discovery and Data Mining Explorations*. 5, 48-55 (2003)