

Extreme Learning Machine for Microarray Cancer Classification

M Yasodha,
Asst Professor, Dept of Computer Science ,
Dr. N.G.P. Arts and Science College, Coimbatore,
Tamil Nadu, India.
Email: vmyasodha@gmail.com

M. Selva Boopathi,
Research Scholar, Dept of Computer Science ,
Dr. N.G.P. Arts and Science College, Coimbatore,
Tamil Nadu, India.
selvaalamelu.cs@gmail.com

Abstract:-Cancer is a diseases in which a set of cells has not able controlled their growth, attack that interrupts upon and destroy the nearest tissues or spreading to other locations in the body. Cancer has become one of the perilous diseases in the present scenario. In this paper, the recently developed Extreme Learning Machine is used for classification problems in cancer diagnosis area. ELM is an available learning algorithm for single layer feed forward neural network. The advanced and developed methodology known for cancer multi classification using ELM microarray gene expression cancer diagnosis, this used for directing multi category classification problems in the cancer diagnosis area. ELM avoids many problems, improper learning rate and over fitting commonly faced by iterative learning methods and completes the training very fast. The performance of classification ELM on three benchmark microarray data for cancer diagnosis, namely Lymphoma data set, Leukemia data set, SRBCT data set. The results of experiments with RVM and ELM shows that for many categories of ELM still outperformer with RVM.

Keywords: Microarray Data, RVM, ELM

1. Introduction

The cancer is one of the terrible disease set found in most of the human being, which is one of the important studies for research in the current century. Cancer is fundamentally described by an abnormal, uncontrolled growth that may destroy and attack other healthy body parts. There are billions of cells in the human body and most of the cells in the cells have an not enough life-span and required cell is capable of duplicating themselves. Millions of cell divisions and replications occur daily in the body and it's amazing that the procedure occurs.

Extreme Learning Machine (ELM) is a currently popular neural network architecture based on random projections. It has one hidden layer with random weights, and an output layer whose weights are determined analytically. Both training and prediction are fast compared with many other non linear methods. ELM, although introduced as a fast method for training a neural network, is in some sense closer to a kernel method in its operation.

In this paper we are going to discuss about the Extreme Learning Machine in detail and also the experiment result of ELM. We compared the experimental result of ELM with RVM(Relevant Vector Machine). ELM perform better than the RVM, in case of accuracy, processing speed and all.

2. Methodology

Microarray analysis is not simple because of the large number of genes, which are investigated concurrently. By incorporating several factors of interest (for instance time and different treatments) in the experimental design, the interpretation of the data becomes even more tricky. The influence of the factors of interest should be separated from each other to draw conclusions from the data analysis. To tackle these problems, a new methodology called ELM is proposed. Nothing like traditional implementations and learning theory, from function approximation point of view, ELM theory shows that the hidden node parameters can be completely independent from the training data.

3. Extreme Learning Machine

A standard single layer feed forward neural network with n hidden neurons and activation function $g(x)$ can be automatically modeled as.

$$\sum_i^n \beta_i g(w_i x_j + b_i) = d_j, j=1, 2, \dots, N$$

Where w_i is the weight vector connecting inputs and the I th hidden neurons β_i is the weight vector connecting the i th hidden neurons and outputs neurons d_j is the output from ELM for data point j.

With N data point in a pair as $(x_j t_j)$, $x_i \in R^m$ and t_j is the corresponding output for data point x_j , the ideal case is training with zero errors, which can be represented as

$$\sum_i^n = 1 \beta_i g(w_i x_j + b_i) = t_j, j=1, 2. . . N$$

The above equations can be written compactly as.

$$H\beta=T$$

Where,

$$H = \begin{bmatrix} g(w_1 x_1 + b_1) & \dots & g(w_n x_1 + b_n) \\ \vdots & \dots & \vdots \\ g(w_1 x_N + b_1) & \dots & g(w_n x_N + b_n) \end{bmatrix}_{N \times n}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_n^T \end{bmatrix}_{m \times m} \text{ and } T = \begin{bmatrix} t_1^T \\ t_N^T \end{bmatrix}_{N \times m}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_n^T \end{bmatrix}_{m \times m} \text{ and } T = \begin{bmatrix} t_1^T \\ t_N^T \end{bmatrix}_{N \times m}$$

So the solution is;

$$\hat{\beta} = H^+T$$

Where H^+ is called Moore-Penrose generalized inverse [7]. The most important properties of this solution as claimed by the authors [7] are:

1. Minimum training error
2. Smallest norm weights and best generalization performance
3. The minimum norm least-square of $H\beta=T$ is unique, which is $\hat{\beta} = H^+T$

So finally, the ELM algorithm is [7]:

Give a training set $\{(x_i, t_i) | x_i \in R^n, t_i \in R^m, i= 1, \dots, N\}$, activation function $g(x)$, and N hidden neurons,

Experiment and Result

The performance of ELM algorithm for Multicategory cancer diagnosis, three benchmark microarray data sets, Lymphoma data set, Leukemia data set SRBCT data set are used in this paper. The ELM performs better.

Lymphoma

Lymphoma is cancer of the lymph system (or lymphatic system), which is part of our immunity. It is characterized by the formation of solid tumors in the immune system.¹ The cancer affects immune cells called lymphocytes, which are white blood cells.

Leukemia

The DNA of immature blood cells, mainly white cells, becomes damaged in some way. This abnormality causes the blood cells to grow and divide chaotically. Normal blood cells die after a while and are replaced by new cells which are produced in the bone marrow. The abnormal blood cells do not die so easily, and accumulate, occupying more and more space. As more and more space is occupied by these faulty blood cells there is less and less space for the normal cells - and the sufferer becomes ill. Quite simply, the bad cells crowd out the good cells in the blood.

SRBCT

The small-, round-, blue-cell tumor (SRBCT), also known as a small-blue-round-cell tumor (SBRCT) or a small-round-cell tumor (SRCT), is any one of a group of malignant neoplasm's that have a characteristic appearance under the microscope, i.e. consisting of small round cells that stain blue on routine H&E stained sections. These tumors are seen more often in children than in adults. They typically represent undifferentiated cells. The predominance of blue staining is due to the fact that the cells consist predominantly of nucleus, thus they have scant cytoplasm.

Testing Accuracy and Execution Time

In this paper going to discuss about compared Accuracy and Execution time for ELM with RVM for following data sets Lymphoma data set, Leukemia data set SRBCT data set. The ELM was better performer than RVM.

Data Set	No. of Gene Combinatio n	Accuracy		Execution Time (Second)	
		EL M	RV M	ELM	RV M
Lymphom a	100,2	78%	69%	23.2 5	50.25
Leukemia	100,3	82%	70%	31.5 4	45.49
SRBCT	100,4	81%	76%	21.2 2	32.88

Table1: Comparison of Accuracy and Execution Time between ELM and

The Table 1 represents the accuracy and execution time for gene expression data using the ELM technique. The comparison of RVM and ELM approaches are evaluated using three datasets lymphoma, leukemia and SRBCT. The ELM is more proficient performance than the existing technique.

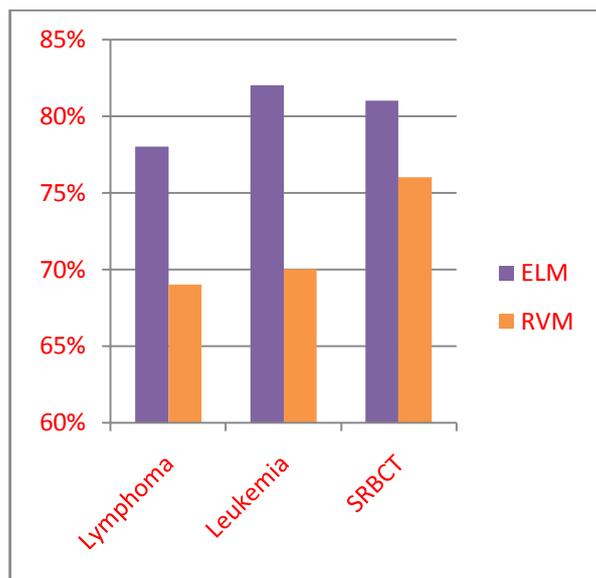


Figure 1: Comparison of Accuracy between ELM and RVM

The accuracy and execution time represents the better outcome using proposed ELM when compared to the RVM techniques. Figure 3 shows the accuracy value for RVM and proposed ELM. Compared to RVM the accuracy value is higher in ELM approach. In Figure 4 represents the execution time for comparable RVM and ELM for gene expression data. In ELM technique execution time is smaller than the RVM.

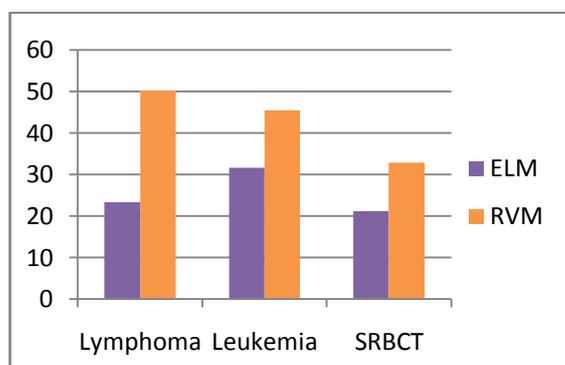


Figure 2: Execution Time between ELM and RVM

4. Conclusion

Cancer research is one of the key research fields in the medical field. Exact prediction of several cancer has higher value in offering enhanced treatment and pain reduction on the patients. Cancer diagnosis problem based on microarray data has become an important field of research area. In this paper, ELM algorithm is used for classification. It performs better than the RVM algorithm. The experiment is performed with the help of lymphoma, leukemia, SRBCT data set. The experimental result shows

that the ELM technique results in better accuracy and consumes less time for classification when compared to the conventional techniques.

REFERENCES

- [1] Kinzler, Kenneth W.; Vogelstein, Bert, "Introduction", the genetic basis of human cancer (2nd, illustrated, revised ed.). New York: McGraw-Hill, Medical Pub. Division, 2002.
- [2] G.-B. Huang and C.-K. Siew, "Extreme Learning Machine: RBF Network Case," Proc. Eighth Int'l Conf. Control, Automation, Robotics, and Vision (ICARCV '04), Dec. 2004
- [3] D. Serre, Matrices: Theory and Applications. Springer-Verlag, 2002.
- [4] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal Approximation Using Incremental Constructive Feedforward Networks with Random Hidden Nodes," IEEE Trans. Neural Networks, vol. 17, no. 4, pp. 879-892, 2006.
- [5] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for Classification of Tumors Using Gene Expression Data," J. Am. Statistical Assoc., vol. 97, no. 457, pp. 77-87, 2002.
- [6] Runxuan Zhang, Guang-Bin Huang, Narasimhan Sundararajan, and P. Saratchandran, " Multicategory Classification Using an Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis, " vol 4,no 3,july-september 2007.
- [7] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science, vol. 270, pp. 467-470, 1995.
- [8] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistical Assoc., vol. 97, pp. 77-87, 2002.
- [9] R. Linder, D. Dew, H. Sudhoff, D. Theegarten, K. Remberger, S.J. Poppl, and M. Wagner, "The 'Subsequent Artificial NeuralNetwork' (SANN) Approach Might Bring More Classificatory Power to ANN-Based DNA Microarray Analyses," Bioinformatics, vol. 20, no. 18, pp. 3544-3552, 2004
- [10] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks," Proc. Int'l Joint Conf. Neural Networks (IJCNN '04), July 2004.
- [11] G.-B. Huang and C.-K. Siew, "Extreme Learning Machine: RBF Network Case," Proc. Eighth Int'l Conf. Control, Automation, Robotics, and Vision (ICARCV '04), Dec. 2004.
- [12] G.-B. Huang and C.-K. Siew, "Extreme Learning Machine with Randomly Assigned RBF Kernels," Int'l J. Information Technology, vol. 11, no. 1, 2005.
- [13] G.-B. Huang, Q.-Y. Zhu, K.Z. Mao, C.-K. Siew, P. Saratchandran, and N. Sundararajan, "Can Threshold

- Networks Be Trained Directly?" IEEE Trans. Circuits and Systems II, vol. 53, no. 3, pp. 187-191, 2006.
- [14] M.-B. Li, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "Fully Complex Extreme Learning Machine," *Neurocomputing*, vol. 68, pp. 306-314, 2005.
- [15] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures," *Proc. Nat'l Academy Sciences, USA*, vol. 98, no. 26, pp. 15149-15154, 2002.
- [16] R. Linder, D. Dew, H. Sudhoff, D. Theegarten, K. Remberger, S.J. Poppl, and M. Wagner, "The 'Subsequent Artificial Neural Network' (SANN) Approach Might Bring More Classificatory Power to ANN-Based DNA Microarray Analyses," *Bioinformatics*, vol. 20, no. 18, pp. 3544-3552, 2004.
- [17] O. Troyanskaya et al., "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.
- [18] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA*, vol. 98, pp. 11 462-11 467, 2001.
- [19] Lipo Wan, Feng Chu and Wei Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 1, Pp. 40-53, 2007.
- [20] Ahmad M. Sarhan, "Cancer Classification based on Microarray Gene Expression Data using DCT and ANN", *Journal of Theoretical and Applied Information Technology*, 2009.
- [21] G. W. Wayt. The unseen genome: beyond DNA. *Scientific American*, 289(6): 106-114 (2003).
- [22] M. Ringner, C. Peterson, and J. Khan, "Analyzing Array Data Using Supervised Methods," *Pharmacogenomics*, vol. 3, no. 3, pp. 403-415, 2002.
- [23] C. Chandrasekar, P.S. Meena /*International Journal of Engineering Research and Applications (IJERA)* ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2, Mar-Apr 2012, pp.229-235