International Journal of Information Processing, 9(3), 1-16, 2015
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Classification of Microarray Data using Extreme Learning Machine Classifier

Mukesh Kumar[a], Sandeep Singh[a], Santanu Kumar Rath[a]

[a]Department of Computer Science and Engineering, National Institute of Technology, Rourkela, Orissa 769008 India,
Contact: mkyadav262@gmail.com, singhit.1507@gmail.com, skrath@nitrkl.ac.in

Microarray dataset often contains a huge number of insignificant and irrelevant features that might lead to loss of useful information. The classes with both high relevance as well as high significance, feature sets are generally preferred for selecting the features, which determine the classification of samples into their respective classes. This property has gained a lot of significance among the researchers and practitioners in DNA micro array classification. In this Paper, Extreme Learning Machine (ELM) classifier, which has gained a lot of popularity due to its very fast learning capability have been considered to classify microarray data sets using t-test as a feature selection method. Further, this paper presents a comparative analysis on the obtained classification accuracy by coupling ELM classifier with different kernel function and well known SVM classifier available in the literature. Performance parameters available in literature such as precision, recall, specificity, F-Measure, ROC curve and accuracy are applied in this comparative analysis to analyze the behavior of the classifiers. From the proposed approach, it is apparent that ELM classifier is the suitable classification model as compare with SVM.

Keywords : Classification, Extreme Learning Machine, Gene selection, Microarray, t-test.

1. INTRODUCTION

Accurate diagnosis of a disease like *cancer* is diagnosis of any disease in particular *cancer*, is vital for successful application of any specific therapy. Even though the classification of cells into cancerous and non-cancerous categories in relation to cancer diagnosis has improved quite significantly over the past few years, still the research is being carried out and there is a scope for improvement in proper diagnosis. This objective can be achieved with the application of less subjective models. Recent development in diagnosis, indicates that DNA microarray provides an insight to cancer classification at gene level. This is due to their capability in measuring abundant messenger ribonucleic acid (mRNA) transcripts for numerous genes concurrently.

Microarray based gene expression profiling has been emerged as an efficient technique for cancer classification as well as for its diagnosis, prognosis, and treatment purposes [1]. In recent years, Deoxyribonucleic acid (DNA) microarray technique has shown a great impact in determining the *informative genes* that cause cancer [2,3]. The major drawback that exists in microarray data is the curse of dimensionality problem [4]. This problem hinders the useful information of data set and leads to computational instability. Therefore, the selection/extraction of relevant features (genes) remains an imperative in the analysis of microarray data of cancer. A good number of feature (gene) extraction techniques and classifiers based on machine learning techniques have been proposed by various researchers and practitioners [5–9].

The main objective of the feature selection (FS) is to (a) avoid over-fitting and improve model (classifier) performance. (b) provide faster and more cost effective models and (c) gain a deeper insight into the underlying processes that generate the data.

Table 9
Average Training, Average Testing Accuracy and CPU Time (in Seconds) of ELM and SVM with Different Kernel Function for Breast Cancer Dataset.

	Linear kernel		Polynomial kernel		RBF kernel		Tansig kernel	
ELM	C=0.03125		$\gamma = 0.75$, $b = 1$, C=0.51563		$\gamma = 0.3125$, C=0.28125		$\gamma = 0.046875$, C=24	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
	80.18(0.0043)	78.35(0.00058)	82.45(0.0058)	78.35(0.00091)	97.72(0.0081)	82.47(0.0018)	76.99(0.0083)	78.35(0.0020)
SVM	C=32		$\gamma = 0.125$, $b = 1$, C=32		$\gamma = 1$, C=4		$\gamma = 0.5$, C=32	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
	82.02	83.44(135)	81.11	80.67(938)	82.71	81.56(185)	83.27	84.44(177)

Table 10
Average Training, Average Testing Accuracy and CPU Time (in Seconds) of ELM and SVM with Different Kernel Function for Ovarian Cancer Dataset.

	Linear kernel		Polynomial kernel		RBF kernel		Tansig kernel	
ELM	C=0.03125		$\gamma = 0.6$, $b = 1$, C=0.03125		$\gamma = 0.03125$, C=0.03125		$\gamma = 0.03125$, C=4.8	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
	100(0.019)	100(0.0028)	99.03(0.054)	99.20(0.0058)	100(0.039)	99.21(0.0059)	99.61(0.043)	99.60(0.0067)
SVM	C=32		$\gamma = 32$, $b = 1$, C=32		$\gamma = 1$, C=32		$\gamma = 0.06255$, C=32	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
	99.52	100(146)	99.43	99.23(171)	99.86	100(720)	99.77	84.44(177)

Hence, from the obtained results, it can be concluded that feature selection plays a significant role in the classification of microarray data, into cancerous and non-cancerous ones.

6. CONCLUSIONS

In this paper, an attempt has been made to design classification models for classifying the samples of microarray data into their respective classes. Hence, a classification framework was designed using ELM classifier with various kernel functions. Feature selection was carried out using t-test. 10-fold CV technique was applied to enhance the performance of the classifiers. The performance of the classifiers for all three data sets were evaluated using performance parameters. From the computed result, it is observed that ELM with RBF function as classifier yields better result when compared with ELM with remaining kernel functions and the existing classifiers available in literature.

Further, the applicability of machine learning techniques such as genetic algorithm (GA), particle swarm optimization (PSO), *etc.*, in combination with ELM can be studied to obtain better classification of microarray data set. This hybridization may help in reducing the complexity of the classification model.

REFERENCES

1. Golub Todd R, Slonim Donna K, Tamayo Pablo, Huard Christine, Gaasenbeek Michelle, Mesirov Jill P, Coller Hilary, Loh Mignon L, Downing James R and Caligiuri Mark A. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *in Journal of science*, 286(5439):531–537, 1999.
2. Leung Yuk Fai and Cavalieri Duccio. Fundamentals of cDNA Microarray Data Analysis, *Elsevier Journal on TRENDS in Genetics*, 19(11):649–659, 2003.
3. Flores M, Hsiao TH, Chiu YC, Chuang EY, Huang Y and Chen Y. Gene Regulation, Modulation and their Applications in Gene Expression Data Analysis, *Advances in Bioinformatics*, pages 11–22, 2013.
4. Lee George, Rodriguez Carlos and Madabhushi Anant. Investigating the Efficacy of Nonlinear Dimensionality Reduction Schemes in Classifying Gene and Protein Expression Studies, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3):368–384, 2008.
5. Lee Kyeong Eun, Sha Naijun, Dougherty Edward R, Vannucci Marina and Mallick Bani K. Gene Selection: A Bayesian Variable Selection Approach, *Bioinformatics, Oxford Univ Press*, 19(1):90–97, 2003.
6. Peng Yanxiong, Li Wenyuan and Liu Ying. A Hybrid Approach for Biomarker Discovery

from Microarray Gene Expression Data for Cancer Classification, *Cancer informatics, Libertas Academica*, 2:301–315, 2006.
7. Wang Lipo, Chu Feng and Xie Wei. Accurate Cancer Classification using Expressions of Very few Genes, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(1):40–53, 200.
8. Deb Kalyanmoy and Raji Reddy A. Reliable Classification of Two-class Cancer Data using Evolutionary Algorithms, *Elsevier Journal on BioSystems*, 72(1):111–129, 2003.
9. Hernandez Jose Crispin Hernandez, Duval Béatrice and Hao Jin-Kao. A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data, *A Book on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, Springer pages 90–101, 2007.
10. Sheskin David J. Handbook of Parametric and Nonparametric Statistical Procedures, crc Press, 2003.
11. Huang Guang-Bin, Wang Dian Hui and Lan Yuan. Extreme Learning Machines: A Survey, *International Journal of Machine Learning and Cybernetics*, Springer, 2(2):107–122, 2011.
12. Schölkopf Bernhard, Smola, Alexander and Müller Klaus-Robert. Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural computation*, MIT Press, 10(5):1299–1319, 1998.
13. Hang Xiyi. Cancer Classification by Sparse Representation using Microarray Gene Expression Data, *IEEE International Conference on Bioinformatics and Biomeidcine Workshops* , pages 174–177, 2008.
14. Ye Jieping, Li Tao, Xiong Tao and Janardan Ravi. Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(4):181–190, 2004.
15. Bharathi A and Natarajan A M. Cancer Classification of Bioinformatics data using ANOVA, *International Journal of Computer Theory and Engineering*, 2(3):369–373, 2010.
16. Salem Dina Ahmed, Seoud Abul, Ahmed Rania and Ali Hesham Arafat. MGS-CM: A Multiple Scoring Gene Selection Technique for Cancer Classification using Microarrays, *International Journal of Computer Applications*, 36(6), 2011.
17. Sun Xin, Liu Yanheng, Xu Mantao, Chen Huiling, Han Jiawei and Wang Kunhao. Feature Selection using Dynamic Weights for Classification, *Knowledge-Based Systems*, Elsevier, 2012.
18. Yeh Wei-Chang, Yeh Yuan-Ming, Chiu Cheng-Wei and Chung Yuk Ying. A Wrapper-Based Combined Recursive Orthogonal Array and Support Vector Machine for Classification and Feature Selection, *Modern Applied Science*, 8(1):11-22, 2013.
19. Jain Yogendra Kumar and Bhandare Santosh Kumar. Min Max Normalization Based Data Perturbation Method for Privacy Protection, *International Journal of Computer and Communication Technology (IJCCT)*, 2(8):45–50, 2011.
20. Mukesh Kumar and Santanu Kumar Rath. Classification of Microarray Data Using Kernel Fuzzy Inference System, *International Scholarly Research Notices*, Hindawi Publishing Corporation, 2014.
21. Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn. Use of Proteomic Patterns in Serum to Identify Ovarian Cancer, *The lancet*, Elsevier, 359(9306):572–577, 2002.
22. Laura J van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, *nature*, Nature Publishing Group, 415(6871):530–536, 2002.
23. Cagatay Catal. Performance Evaluation Metrics for Software Fault Prediction Studies, *Acta Polytechnica Hungarica*, 9(4):193–206, 2012.

Mukesh Kumar has completed master's degree in Computer Science and Engineering with specialization of Software Engineering from NIT Rourkela, India in 2013. He is currently a research scholar in Computer Science at at NIT Rourkela, India. His areas of interest are mostly on Software Engineering, Bioinformatics, Big Data Analytics and Machine Learning.

Sandeep Singh is pursuing his masters degree in Computer Science and Engineering with specialization of Software Engineering from NIT Rourkela, India. His areas of interest are mostly on software Engineering, Bioinformatics and Machine Learning.

Dr. Santanu Kumar Rath is a Professor in the Department of Computer Science and Engineering, NIT Rourkela since 1988. His research interests are in Software Engineering, System Engineering, Bioinformatics and Management. He is a Senior Member of the IEEE, USA and ACM, USA and Petri Net Society, Germany.