Machine Learning Techniques for Cancer Risk Prediction

Bichitrananda Patra[1], Santosini Bhutia[2], Niranjan Panda[3]
[1,3]Department of Computer Science and Engineering, Siksha O Anusandhan University,
Bhubaneswar, Odisha
[1]bichitranandapatra@soa.ac.in, [2]santosini.bhutia@gmail.com, [3]niranjanpanda@soa.ac.in

Article Info
Volume 83
Page Number: 7414 - 7420
Publication Issue:
May-June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *18 May 2020*

***Abstract*:**
Microarray data analysis plays a vital role in cancer classification and diagnosis. But it's a big challenge to achieve high level of accuracy in cancer classification with large set of genes. The totalcounts of features areregularlygreater than that of the number of instances. For this reason, it needs to achieve feature selections for organisation of genes. Feature collection lowers theproblematicthrough selecting informative features from datasets. In this study, four feature selection methods with ranker search techniques tool of Weka are used to select top 100 informative genes, the classification technique Support Vector Machine, Random Forest, Random Tree are applied to these selected genes to conduct experimental work on the presented data-sets. The experimentally result that the projected feature assortment and lengthfall in data volume stretchesimproved result of accuracy.

Keywords: Cancer Classification, Feature Selection, Microarray data, SVM, Random Forest, Random Tree

1 Introduction

In entire lifespan, healthy cells replicate and replace the ageing cells in a controlled process. Cancer is a disease that causes cells to multiply in an uncontrolled manner, leading to tumour and weakening the immune system. In medical science it is a major research area to provide better cure. In the world the latest technology is Microarray data analysis, which helps for better treatment of cancer classification and diagnosis from the vast number of genes. But microarray datasets are expensive and complicated in nature, hence it require careful experimentation and appropriate tools for producing good result. The microarray technology has been used for monitoring of genes and tumour classifybygenetic factorcountenance data [1,2,3,4]. However, during experimental process, large volumes of data are generated with errors which pose big challenge to analyse genetic factormanifestation data.

Microarray [5] is identified as genetic factor or DNA, which is used towards represent genetic factorcountenance labels for a huge quantity of gene. It is capable of producing grouping and expectation of dissimilarcategories of tumourdata. However, these huge datasets are often random. Hence researchers have used various optimization techniques for most advantageous feature selection for cancer classification. In this study, we principally focus on gene section or feature section, which is a process of selecting a subset of significant features for constructing a model for cancer classification. Feature selection is also known as variable selection. Feature selection techniques are applicable when there are many features with analogously few samples, like DNA microarray data.

2 Methods

There are different classification's algorithms used in this experimental study. In Weka 3.9.3 the classifiers are categorized into different groups. These are Bayes, Functions based, Lazy based, Rule based, Tree based classifiers etc. Different optimal algorithms have been chosen in the designing of effective and accurate models by reducing the optimal datasets. In this paper it includes Support Vector Machine, Random Forest and Random Tree. These algorithms are described in the following section.

2.1 Support Vector Machine

Support Vector Machine [6] is a powerful method of machine learning technology for both classification and regression. SVM[7] are supervised learning methods, performs classification by constructing a hyper plane. This hyper plane gives separation line which shows the maximum distance to the nearest training points of other class, hence it is called functional margin. The vectors which define this hyper plane are called support vectors and the algorithm which defines optimal hyper plane gives maximum margin. Here fewer numbers of trained samples isrecycled to becomeutmostoptimumhyper plane which result high classification accurateness. SVM uses kernel function to high margin classification and hence proved to be powerful classifier.

SVM supports C-SVC, nu-SVC, one-class SVM for classification and epsilon-SVR, nu-SVR for regression, and also supports different kernel types: linear, poly, rbf, sigmoid. For microarray data classification, in this paper, lib-SVM [8] is considered with C-SVC SVM type and linear kernel type.

2.2 Random Forest

Random Forest [9] is one of the most powerful classes of supervised machine learning algorithm. It accomplished of solving together classification and regression problem. Itgenerates the junglethrough the number of decision trees. The further tree in the forest, the further accuracy in the calculation and less error is generated. To categorize anoriginal object based on attribute, every tree providesorganisation and this tree votes used for that class. The forest chose the classification having the most votes of all the other trees in the forest.

When a large proportion of data are missing, Random Forest can handle missing values and maintain accuracy. Random Forest does not over fit when more number of trees is added to the model. Random forest has the control to handle bulky dataset withcomplex dimensionality.

2.3 Random Tree

Random Tree [19] is a supervised classifier which produces many individual learners. Random Tree is introduced by Leo Breiman and Adele Cutler. This algorithm is capable of solving both classification and regression problem. To construct a decision tree it applies the bagging idea with a random set of data. In standard tree every node id divided by the best possible division amongstthe entirely variables. But in case of random forest, every node is divided by the finest among the subgroup of predicators arbitrarilyselected at that node. The mechanisms of Random Tree classifier are taking the feature vector as input, classify it through every tree in the forests and lastly produce the output class layer that established the mainstream of votes.

3 Experimental Results and Discussion

3.1 Datasets Used

ALL-AML Based Leukaemia: Dataset consist of 72 samples from which 25 AML

(Acute Myeloid Leukemia) and 47 ALL (Acute Lymphoblastic Leukemia). Every sample comprises 5148 gene appearance levels.

Lung cancer: Dataset consists of 203 samples. The tumor response to neoadjuvant treatment was assessed after the samples, which (samples) were obtained before treatment. Everysamples contain 12601 gene appearance levels.

Lymphoma cancer: Dataset consists of 77 samples from which 58 samples are Diffuse Large B-cell Lymphoma (DLBCL), 19 samples are Follicular Lymphoma (FL). Two B-cell lineage malignancies Follicular Lymphoma (FL) and Diffuse Large B-cell Lymphoma (DLBCL) have very different clinical presentations, response to therapy and natural histories. However, Follicular lymphoma develop gradually over time in order to obtain the clinical and morphologic property of DLBCLs, some subsets of Diffuse large B-cell lymphoma have chromosomal translocations characteristic of FLs. Each sample contains 7030 gene expression.

SRBCT: Dataset consist of 83 samples. The Small Round Blue Cell Tumor (SRBCT) isconsisting of four dissimilarinfant tumors namedsince of their related presence on predictable histology. Still, perfectanalysis is required since the treatment possibilities, comebacks to therapy and prognoses differgenerallycontingent on the analysis. Each sample contains 2309 gene expression.

Brain Tumor: The dataset consists of 90 samples which contain five types of tumors (i.e., Astrocytma,Meningioma, Oligodendroglioma,Medulloblastoma and Ependymomatumor). Each sample contains 5921 gene expression.
Here we display the observed results of classifierby different feature selections such as Correlations, Gain Ratio, Information Gain andReliefF models. In this concept we have experimented by three different classifiers such as SVM, Random Forestand Random Tree respectively for the given binary and multi class datasets.

3.2 Analysis

Here we are presenting our analytical result and discussing about its efficiency on microarray genes expressions. There are four different feature selection approaches that are Correlation, Gain Ratio, Information Gain, ReliefF with ranker is used and then top 100 informative genes are selected. Finally the selected genes are classified by these three classifiers: SVM, Random Forest and Random Tree. Various microarray datasets from cancer gene expression studies are available publicly. Among them five datasets (Leukaemia, Lung, Lymphoma, SRBCT, Brain tumour) are used in this paper and the accuracy is estimated.

Table 1 Accuracy (%) for libSVMClassifier

Dataset	Correlation	GainRatio	IG	RelifF
Leukemia	98.61	97.22	95.83	97.22
Lung	95.07	93.10	96.05	95.07
Lymphoma	100	96.10	97.40	97.40
SRBCT	100	100	100	100
Brain Tumor	91.11	86.66	90	88.88

Table2 Accuracy (%) for Random Forest Classifier

Dataset	Correlation	GainRatio	IG	ReliffF
Leukemia	98.61	98.61	98.61	98.61
Lung	92.61	87.68	94.58	93.10
Lymphoma	94.80	93.50	93.50	94.80
SRBCT	100	100	98.79	100
Brain Tumor	84.44	85.55	90	85.55

Table 3 Accuracy (%) for Random Tree Classifier

Dataset	Correlation	GainRatio	IG	ReliffF
Leukemia	94.44	90.27	90.27	84.72
Lung	83.74	74.87	85.71	85.71
Lymphoma	77.92	81.81	96.10	85.71
SRBCT	83.13	86.74	92.77	84.33
Brain Tumor	75.55	72.22	85.55	73.33

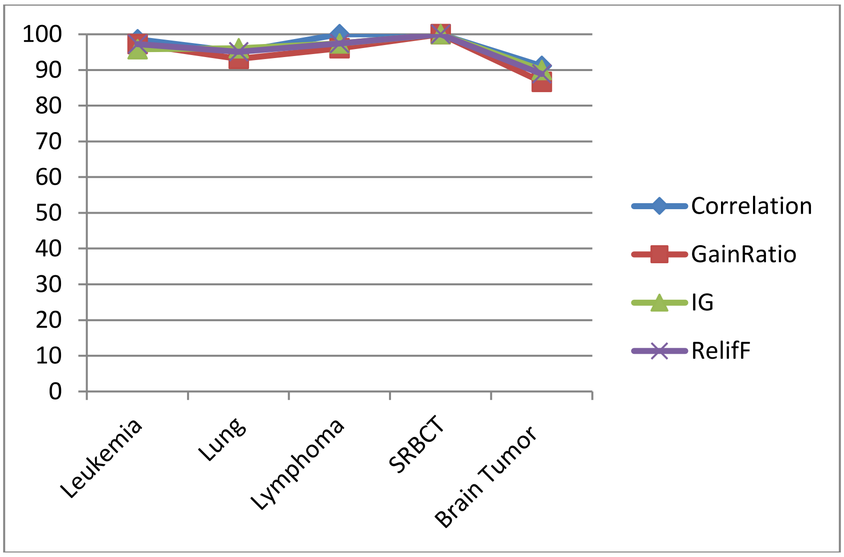

Fig 1. Accuracy Percentage of SVM Classifier

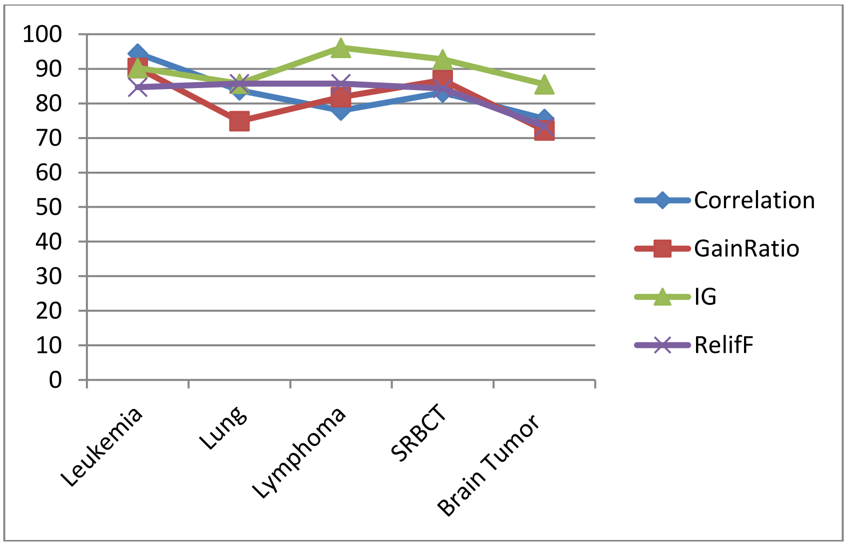

Fig 2. Accuracy Percentage of Random Forest Classifier

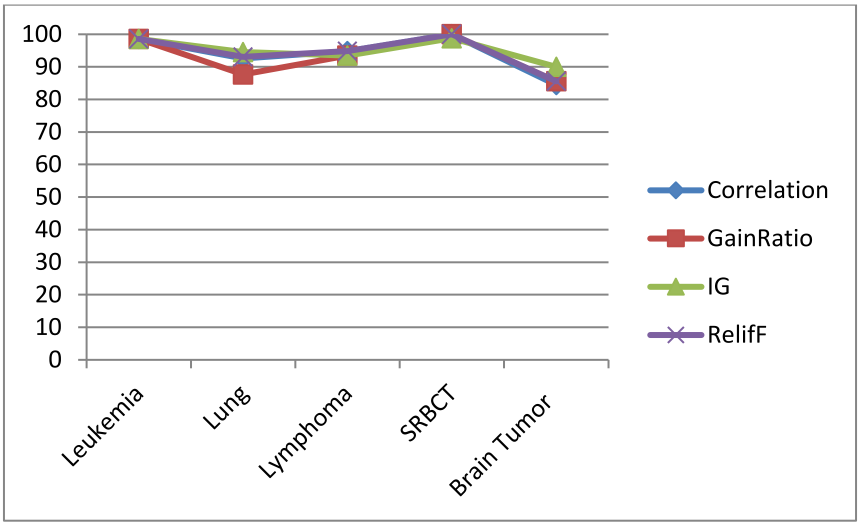

Fig 3. Accuracy Percentage of Random Tree Classifier

Fig 1demonstrations the satisfactory accurateness of the classifiers of all elementthen it is related over the accuracy originatesince additional predictable model represented on Table 1, Table 2 and Table 3. Limitations are setting for NNSVM classifiers are usual as pertracks: Linear SVM classifications use C-SVC as linear classifier, 100 is batch size, linear as kernel type and0.5 for nu. Small variations are approaching while they are usedas self-classifier.

The comparative study is represented in Table 4. This comparison surmises that the projected method performs analogous accurateness for all the microarray dataset.

Table 4 Percentage Accuracy of specificcurrent Researchers

Author	Dataset	Methods		Accuracy
		Feature	Classifier	
Vural&Subasi(2015)	Leukemia	SVD &IG	SVM	97.14
	Lung			93.60
	Lymphoma			98.70
	SRBCT			95.18
Vural&Subasi(2015)	Leukemia	SVD &IG	RF	91.43
	Lung			90.64
	Lymphoma			90.91
	SRBCT			86.75
P Mohapatra& S Chakravarty(2015)	Leukemia	MPSO	SVM KNN NB	93.85 91.12 94.58
M Panda(2017)	SRBCT	FFS	DL	93.98
	Lung			93.11
	DLBCL			89.36
M Panda(2017)	SRBCT	ES	DL	83.14
	Lung			94.1
	DLBCL			91.49

Vural and Subasiproposed an innovative research on “Data-Mining Techniques to Classify Microarray Gene Expression Data Using Gene Selection by SVD and Information Gain”[10]. In this study, Singular Value Decomposition (SVD) has been recycled to choicerevealing genes and to decrease the inessential evidence. Moreover, Information Gain (IG) is recycled to control useful structures and finally SVM, ANN, Random Forest classification techniques are applied for better classification result.

PMohapatra& S Chakravarty proposed “Modified PSO based Feature Selection for Microarray Data Classification”[11]. At present paper, Modified Particle Swarm Optimization (MPSO) is recycled to choice important genes from the publicly available biomedical microarray datasets. Furthermore, Support Vector Machines (SVM), Naive Bayesian (NB) and k-Nearest Neighbor (KNN) classifiers are recycled for classification and it is observed that SVM performs better than other two classifiers.

M Panda proposed a research on “Elephant Search with Deep Learning for Microarray Data Analysis”[12]. In this experimental study, the performance of microarray gene expression profiling is demonstrated with Firefly search (FFS) and Elephant Search (ES) based on Deep learning. Deep learning performs better for almost in all datasets. But Deep learning fails to perform good when a convoluted model is select to study from an easy problem.

B N Patra and S. Bisoyi proposed a study on “CFSES optimization Feature Selection with neural network classification for microarray data analysis”[13]. In this untried study, the presentation of microarray gene expression summarizing is established with Elephant Search (ES) based on neural network learning. Neural Network classifiers perform better results in two given datasets. But it is not performedfor multi class datasets.

4 Conclusion

Classification problems have been largely premeditated by researchers in the field of machine Learning. For the last few year researchers have started exploring cancer classification using gene expression. Microarray data can be used in the discovery and prediction of cancer classes. Various approaches are used for efficient gene selection for producing cancer classification results. In this research work, the feature selection methods: Correlation, Gain Ratio, Information Gain (IG), RelifF are used with ranker to rank the attributes with their individual evaluation, and then first 100 informative genes are selected. Finally the selected genes are classified by these three classifiers: libSVM, Random Forest, Random Tree. From the experiment it is observed that libSVM performs better followed by Random Forest and Random Tree.

References:

1. Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.
2. D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander. Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology (RECOMB), 2000, pages 263–272.
3. Lakhani, Sunil R., and Alan Ashworth. "Microarray and histopathological analysis of tumours: the future and the past?." Nature Reviews Cancer 1.2 (2001): 151-157.
4. Nguyen, Danh V., and David M. Rocke. "Classification of acute leukemia based on DNA microarray gene expressions using partial least

squares." Methods of Microarray Data Analysis. Springer US, 2002.109-124.

5. Harrington, Christina A., CarstenRosenow, and Jacques Retief. "Monitoring gene expression using DNA microarrays." Current opinion in Microbiology 3.3 (2000): 285-291.
6. Jaison, B., A. Chilambuchelvan, and A. Kannan. "A Discrete Wavelet based feature extraction and Hybrid Classification technique for Microarray data analysis."
7. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003).
8. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011): 27.
9. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32
10. H. Vural and A. Subasi, Data Mining techniques to classify Microarray geneexpression data using selection by SVD and information gain, Modeling of Artificialinteligence, vol. 6, no-2 pp. 171-182, 2015.
11. P Mohapatra& S Chakravarty, Modified PSO based Feature Selection for Microarray Data Classification IEEE Power, Communication and Information Technology Conference (PCITC)2015
12. Panda, M. Elephant search optimization combined with deep neural network for microarray data analysis. Journal of King Saud University – Computer and Information Sciences (2017)
13. Patra, B.N. &Bisyoi S.K., “CFSES Optimization Feature Selection with Neural Network Classification for Microarray Data Analysis”, published in IEEEXplore2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018 on 21-23 Sept. (2018). Page(s) : 45-50, ISBN-13: 978-1-5386-8431-3.