ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 1, January 2017

Mining of Epigenetic Data for the Effective Prediction of Bladder Cancer

Mohammed Siyad B
Department Computer Science and Engineering
TKM College of Engineering, Kollam, Kerala, India

Abstract: **-- Epigenetic alterations have been associated with a wide variety of diseases including cancer. Bladder cancer is the fourth most common cancer and the ninth driving reason of cancer death. A lot of tools and protocols have been developed for the diagnosis of bladder cancer over the past 5 to 10 years. In this paper, a machine learning approach is proposed for effectively predicting the disease from epigenetic information in the context of bladder cancer. Three different feature selection methods were assessed in combination with three classification methods, using 10-fold cross-validation on the training data set. A model consisting of 151 genes(treated as features) selected through genetic algorithm and random forest classification is identified as the best model with AUC=0.96 from 10-fold cross validation. Most of the selected genes which formed the basis of prediction were allegedly reported in the pathways related to bladder cancer. Hence the best selected model can be effectively applied for better disease diagnosis and prognosis.**

***Index Terms*—Bladder Cancer, Disease Prediction, Epigenetics, Genetic Algorithm.**

I. INTRODUCTION

Bladder cancer is the commonest malignancy of the urinary tract, with the incidence being four times higher in men than in women [1]. The bladder is a hollow organ in the pelvis with flexible, muscular walls. Its main function is to store urine before it leaves the body. Bladder cancer starts when cells in the urinary bladder begin to become wildly. As more malignancy cells create, they can shape a tumor and spread to different regions of the body. If bladder cancer spreads, it often goes first to distant lymph nodes, the bones, the lungs, or the liver [2]. Cigarette smoking, urinary tract infections, occupational exposure to aromatic amines and polycyclic aromatic hydrocarbons, and drugs are risk factors for the disease [3].

Epigenetics is commonly defined as the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in the DNA sequence" [4]. It is one of rapidly growing areas in the field of computational biology. This is mainly due to the ramarkable technological advances which enable genome-scale analysis of epigenetic mechanisms. The different types of epigenetic mechanisms are DNA methylation, Histone modifications and RNA-associated gene silencing. The packaging of DNA inside the nucleus directly influences gene expression and hence any epigenetic modifications can affect gene expression.

Currently, DNA methylation is one of the most broadly studied and well-characterized epigenetic modifications. DNA methylation consists of the addition of a methyl group to carbon 5 of the cytosine within the dinucleotide CpG. DNA Methylation can modify the gene expression. Increased DNA methylation or hypermethylation initiates silencing of tumor suppressor genes and a massive loss of DNA methylation or hypomethylation can activate oncogenes and initiate chromosome instability. DNA methylation is carried out by a family of enzymes called DNA methyltransferases (DNMTs). Most probably, DNA methylation is associated with loss of gene expression. DNA methylation highly correlates to the regulation of gene expression. For gene transcription to happen, the gene promoter region should be easily accessible to transcription factors [5]. The DNA methylation directly prevents the transcription factors from accessing the promoter region which leads to the silencing of genes.

Histone modification is a covalent post-translational modification to histone proteins. Histones are the core of nucleosomes that DNA sequences wrap around [6]. All histones are subject to some level of methylation, acetylation or phosphorylation which would affect the local chromatin structures to enable or repress gene expression. Histone modifications play multifaceted roles for several cellular processes including gene transcription, DNA repair, recombination

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

and DNA replication. The deregulation of this process is implicated in human malignancies.

Noncoding microRNAs play essential role in the maintenance of the gene transcription state through multiple cell divisions. The miRNA seems, by all accounts, to be able to act as either tumor suppressors or oncogenes by influencing distinct genes involved in critical biological processes such as proliferation and differentiation [7]. Several recent studies indicated that miRNA profiles significantly differ between cancer and normal tissues. This distinction in miRNA profiling can classify cancers according to the developmental lineage and differentiation status which lend miRNAs as useful tools in cancer diagnostics and prognosis.

The rest of this paper is organized as follows. Section II describes the motivation of the study. Section III gives the landscape of the methodology adopted in the work. The experimental setup of the proposed work is described in section IV. Section V discusses the results obtained. Section VI gives the concluding remarks with some feature prospects of the work.

II. MOTIVATION

Although a number of intergrated data analysis tools are available and a lot of methylation data analysis protocols were developed, a significant model that predict the disease from epigenetic data is currently in their infancy. Also several new techniques and developments have been introduced in recent years to improve the diagnosis and management of bladder cancer [1]. Here a computational model based on machine learning approach is proposed for the efficient recognition between bladder cancer and reference(or healthy) samples.

III. METHODOLOGY

In this paper, a machine learning approach is proposed for the efficient recognition between bladder cancer and reference(or healthy) samples. Here three feature selection methods were evaluated in combination with three classification methods, using 10-fold cross-validation on the training data set.

A. Applied Feature Selection Methods

Feature selection is one of the dimensionality reduction approaches which identifies a small subset of features that minimize redundancy and maximize relevance to the target [8]. The selected features are capable of discriminating samples that belong to different classes. In this paper, the following feature selection approaches are used.

1) SVM Attribute Evaluation: SVMs have been seriously contemplated and benchmarked against a variety of strategies recently. They are presently one of the best-known classification methods with computational advantages over their competitors. Linear SVMs are particular linear discriminant classifiers. SVMs lend themselves especially well to the analysis of broad patterns. They integrate pattern selection and feature selection in a single consistent framework. SVM method of Recursive Feature Elimination(RFE) is much more robust to data overfitting than different techniques, including combinatorial search. SVM RFE is an application of RFE using the weight magnitude as ranking criterion [9]. This technique assesses the worth of an attribute by utilizing a SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class seperately using a one-vs-all method and then 'dealing' from the top of each pile to give a final ranking [9]. SVM feature selection fundamentally relies on upon having clean information since the outliers play an essential role.

2) Relief: ReliefF is a straightforward and gainful strategy to gauge the nature of qualities of attributes with high trait conditions. The attributes are positioned by the highest correlation with the observed class while considering the separations between various classes. ReliefF searches for k nearest neighbors of a randomly selected instance Im from the same class L (called nearest hits H) and also from each of the different classes (called nearest misses M). A quality estimation Wi for ith attribute is defined and is updated (incremented or decremented) on the values of Im, H and M. After n iterations, all the hits and misses contributions are averaged [10].

3) Genetic Algorithm: A typical GA is an evolutionary process wherein a population of solutions advances over

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

a sequence of generations. Each individual in the population (known as genome or chromosome) represents a candidate solution to the problem. The potential solutions compete and mate with each other to deliver progressively fitter individuals over subsequent generations of solutions. During the reproduction of the next generation, selected individuals are transformed using crossover or mutation processes under a certain crossover probability pc , and mutation probability pm.
The genome (possible set of features) is represented as a binary string of length N that represents the presence or absence of each of the N possible features. A '1' in the string means that feature is selected for process and a '0' represents the absence of that feature. The individuals were assessed for fitness using a fitness function. The individuals with highest fitness were passed onto the next generation. The operation was repeated in each generation. The genome with the highest fitness after a number of generations represents the best feature set [11].

B. Applied Classification Techniques

Classification is the problem of identifying the categories (sub-populations) of a new observation, based on a set of a training data containing observations (or instances) with known categories [8]. Basically, it is a mapping of the feature represented data to a set of labels. The following three classification methods are applied here.

1) J48: J48 Decision tree classifier is an extension of ID3(Iterative Dichotomiser3) developed by the WEKA project team. ID3 constructs a decision tree from a fixed set of examples and is used to classify future samples using information gain of the attributes. In the training phase it identifies the attribute with the highest information gain which most discriminates the various instances clearly. If there is any value of the feature for which the data instances falling within its category have the same value for the target variable, then terminates the branch and assigns with the target value that we have obtained. The process is continued for other attributes until we either get a clear decision or we run out of attributes. In the event that run out of attributes, the corresponding branch will be assigned with a target value that the majority of the items under this branch possess. By checking all the respective attributes and their values in the decision tree model in the order of attribute selection, we can assign or predict the target value of the new instance [12]. In fact in several cases, it was seen that J48 Decision Trees had a higher accuracy than either Naive Bayes or Support Vector Machines [13].

2) Zero R: ZeroR is one of the simplest classification methods which depends on the target and disregards all predictors [14]. Basically this method identifies the majority category. It is useful for determining the baseline performance as a benchmark for other classification methods. It constructs a frequency table for the target and select the most frequent value. Zero R produces the mean for a numeric class or the mode for a nominal class.

3) Random Forest: The Breiman Random Forest is an expansive accumulation of decorrelated decision trees and letting them vote for the most popular class [15]. When used for classification, a random forest obtains a class vote from each tree, and then classifies using majority vote. N subsets are generated at random from the training data. For each subset, a decision tree is created. When an unknown sample arrives, each tree makes a prediction (vote) for that sample. Finally, the class which have the majority votes obtained is selected as the actual class for the sample. Non-linear classification methods (Random Forest and J48) usually perform well compared to linear classifiers [6].

IV. EXPERIMENTAL SETUP

A. Data Set

The experiments were conducted on bladder cancer data downloaded from GEO (NCBI) repository [16], [17]. The DNA methylation and gene expression profiles were obtained separately. Methylation patterns were assayed using the genome-wide Illumina Infinium Human Methylation27 Bead- Chip array. The DNA methylation levels(value) of particular vCpGs range from 0 to 1. A '0' value indicates completely unmethylated and '1' for completely methylated. The data is extracted from 24 samples. The DNA methylation data and gene expression profiles have been under gone various preprocessing operations. The common genes and their expressions in the DNA methylation and gene expression profiles were identified. Log-fold changes in the gene expression levels across the Normal and Disease samples were measured. The genes with similar

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

expression values in both classes do not have a critical role in discriminating the samples. So they were excluded in such a way that the ratio of the means of the expressions in both classes (Normal and Disease) was above 0.98. Thus the total number of genes were reduced to 1748. The WEKA 3.6.13 data mining software is used for feature selection and classification.

B. Feature Selection and Classification

In this paper, three feature selection methods were evaluated in combination with three classification approaches as described in sections III-A and III-B. The training data set underwent 10-fold cross validation on various combinations of feature selection and classification methods, in order to obtain the best model. The Genetic Algorithm has used a population size of 20, a mutation rate of 0.033 and a crossover rate of 0.6. For evaluation the fitness of each chromosome (candidate set), the accuracy of a Random Forest classifier was applied and 20 generations were analyzed. The genes were also evaluated using SVM and ReliefF approaches in association with a Ranker search method. All these selected features were applied to the three classification techniques.

V. RESULTS AND DISCUSSIONS

All the experiments were conducted on bladder cancer data. The preprocessed data (section IV-A) have been applied to three feature selection methods (Support vector Machine- SVM, ReliefF and Genetic Search) in combination with three classification techniques (J48, Random Forest and ZeroR) using 10-fold cross-validation.Total number of instances considered were 24. The results of the classification approaches are summarized in Table I.

Classification	Feature Selection	Correctly Classified	Accuracy	AUC
J48	SVM	18	75%	0.72
	Genetic Search	21	87.5%	0.86
	ReliefF	19	79.16%	0.75
Random Forest	SVM	21	87.5%	0.94
	Genetic Search	22	91.6%	0.96
	ReliefF	22	91.6%	0.94
ZeroR	SVM	18	75%	0.241
	Genetic Search	18	75%	0.241
	ReliefF	18	75%	0.241

Table I
Result Summary of Various Classifiers

From Table1 it is clear that the model with Genetic Search feature selection and Random Forest classification gives the best AUC of 0.96 and it is selected as the best model. It discriminates the labels with an accuracy of 91.6%. The genetic algorithm produced the offsprings based on accuracy of a Random Forest classifier itself as the fitness function. So it always produces the most relevant features to the successive generations and finally 22 instances out of 24 were classified correctly.

ROC analysis is a useful tool for evaluating the accuracy of a statistical model, more specifically for evaluating the performance of diagnostic tests [18], [19]. It is a plot of the true positive rate(TPR) against the false positive rate(FPR) for the different possible threshold points of a diagnostic test. Typically, the true positive rate is on the vertical axis and the false positive rate is on the horizontal axis. ROC curve is a measure of test accuracy through the area under the curve (AUC). The ROC curves of the classification models with the three feature selection techniques are shown in Fig 1,2 and 3. The graphical representation of the performance comparison of the models with various feature selection and classification approaches are given in Fig 4. From these figures it is clear that the model with Genetic Search feature selection and Random Forest classification gives the best AUC of 0.96.

Sensitivity and specificity [20] are two important performance measures which assess more than just a count of correct classifications. Sensitivity(true positive rate, TPR) is the ability of the system to detect disease in a population of diseased individuals. It indicates the percentage of positives that are correctly classified as such. On the other hand, specificity(true negative rate, TNR) is the ability of the system to correctly identify the healthy samples in a healthy population. It represents the correctly classified negatives.

$$Sensitivity = \frac{TP}{(TP + FN)} \qquad (1)$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 1, Issue 1, January 2017

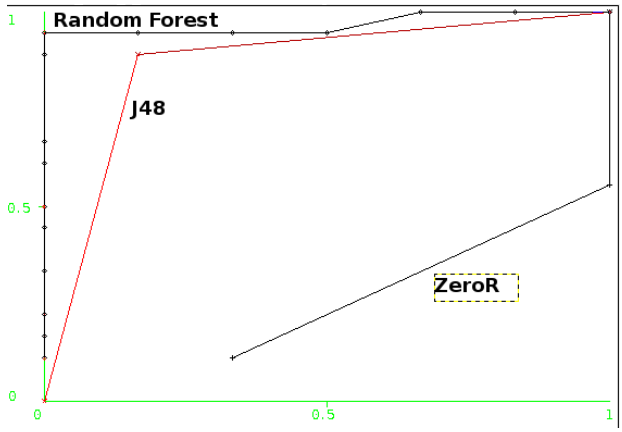

Fig. 1. ROC comparison- Genetic Algorithm approach

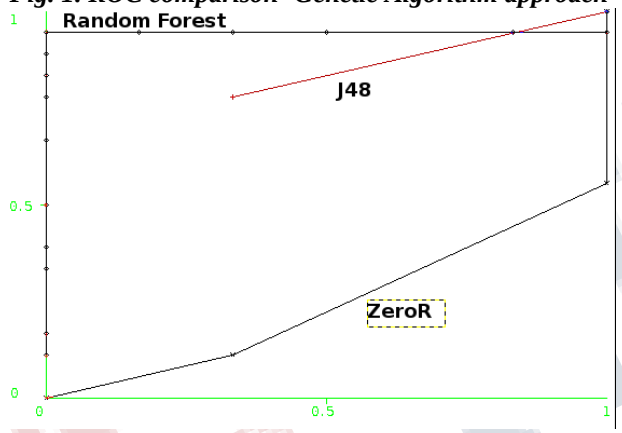

Fig. 2. ROC comparison- SVM approach

$$Specificity = \frac{TN}{(TN + FP)} = 1 - FPR \qquad (2)$$

where, TP is the number of true positives, TN is the number of true negatives, FP, the number of false positives and FN, the number of true negatives in the confusion matrix. FPR indicates the false positive rate. The sensitivity and specificity of the various combinations of feature selection and classification methods are give in Table II. From Table II, it is clear that the

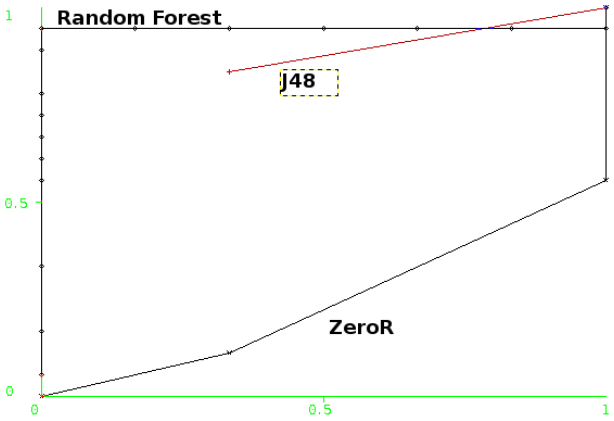

Fig. 3. ROC comparison- ReliefF approach

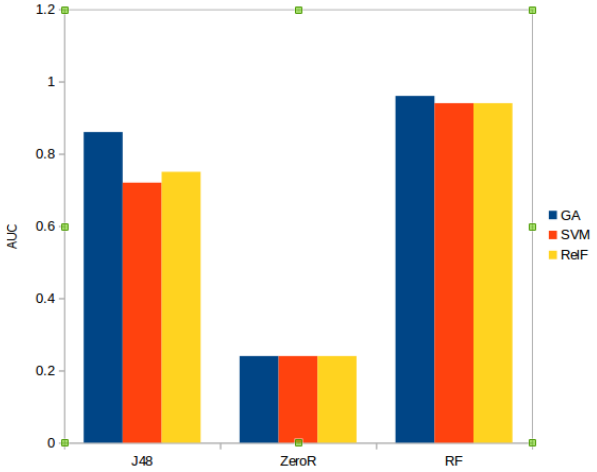

Fig. 4. Performance comparison of models with various feature selection and classification methods

Classification	Feature Selection	Sensitivity	Specificity
J48	SVM	0.77	0.66
	Genetic Search	0.88	0.83
	ReliefF	0.83	0.66
Random Forest	SVM	0.94	0.66
	Genetic Search	0.94	0.83
	ReliefF	0.94	0.83
ZeroR	SVM	1	0
	Genetic Search	1	0
	ReliefF	1	0

Table II
Sensitivity and Specificity of Various Combinations

best model selected (the model with Genetic Search feature selection and Random Forest classification) gives

a sensitivity of 0.94 and specificity of 0.83. These measures reveals that 94% of the diseased people were identified correctly, but 17% of the normal people were incorrectly predicted as having disease. The model with SVM feature selection and Random Forest classification produces a specificity of 0.66 means that 33% of the healthy people were incorrectly predicted as having the disease. But the model with ZeroR classification gives 0 specificity means all healthy samples were incorrectly predicted as diseased, which is dangerous. Thus the specificity and sensitivity are two significant measures of the performance.

The final set of genes selected by the best model were analyzed with the pathways related to bladder cancer in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database for the effective analysis of gene functions [21]. Most of the biological functions like calcium ion transmembrane transport, ATP biosynthetic process, signal transduction etc. of the selected genes (ATP2A1, ATP2B1, MAP2K3, EDNRB, IL2RG etc.) were professedly reported in the biological pathways related to bladder cancer. These results also uncover the biological legitimacy of the proposed best model.

VI. CONCLUSION

Epigenetic changes regulate gene expression and are identified to cause gene expression changes in wide variety of diseases including cancer. In this paper, a data mining approach is proposed to effectively predict a disease using the most relevant features associated with it. The experiments were performed on epigenetic bladder cancer data. Three feature selection methods were applied in combination with three classification methods using 10-fold cross validation on the training data. The model comprised of random forest classification with genetic algorithm based feature selection is selected as the best model with an AUC of 0.96. Most of the selected genes were reputedly reported in the biological pathways related to bladder cancer. In future, those pathways can go about as potential bio-markers for focused medication disclosure and therapeutics.

REFERENCES

[1] G. Cheung, A. Sahai, M. Billia, P. Dasgupta, and M. S. Khan, "Recent advances in the diagnosis and treatment of bladder cancer," BMC Medicine, vol. 11, no. 13, March 2013.

[2] "What is bladder-cancer," http://www.cancer.org/cancer/bladdercancer/ detailedguide/bladder-cancer-what-is-bladder-cancer.

[3] C. Piccinni, D. Motola, G. Marchesini, and E. Poluzzi, "Assessing the association of pioglitazone use and bladder cancer through drug adverse event reporting," Diabetes Care, vol. 34, pp. 1369–1371, 2011.

[4] N. U. Nair, "Computational problems in epigenetics," EDIC Research Proposal, 2010.

[5] L. DHK and M. ER, "DNA methylation: a form of epigenetic control of gene expression," The Obstetrician and Gynaecologist, vol. 12, no. 1, pp. 37–42, January 2010.

[6] J. Li, T. Ching, S. Huang, and L. X. Garmire, "Using epigenomics data to predict gene expression in lung cancer," BMC Bioinformatics, vol. 16, no. S5, pp. 1471–2105, March 2015.

[7] Z. Herce and P. Hainaut, "Genetic and epigenetic alterations as biomarkers for cancer detection," Molecular On Cology1, pp. 26–41, 2007.

[8] J. Tang, S. Alelyani, and H. Liu, Data Classification Algorithms and Applications, Chapter 2- Feature Selection for Classification: A Review. Chapman and Hall/CRC 2014, 2014.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, vol. 46, pp. 389–422, 2002.

[10] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," Machine Learning, vol. 53, pp. 23–69, October 2003.

[11] S. Osowski, T. Markiewicz, and K. Siwek, "Application of support vector machine and genetic algorithm for improved blood cell recognition," IEEE Transactions On Instrumentation And Measurement, vol. 58, pp. 2159–2168, October 2009.

[12] A. K. Yadav and S. Chandel, "Solar energy potential assessment of western himalayan indian state of himachal pradesh using j48 algorithm of weka in ann based prediction model," Renewable Energy, Elsevier, vol. 75, p. 675693, March 2015.

[13] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," IJARCSSE, vol. 3, no. 6, pp. 1114–1119, June 2013.

[14] Surabhi and S. K. Pandey, "Performance evaluation of supervised classification algorithms using data mining," IJSAE, vol. 2, no. 8, pp. 1476–1482, August 2014.

[15] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[16] "NCBI data base," http:///www.ncbi.nlm.nih.gov/geo/query/acc.cgi?accn =gse37816, series GSE37816.

[17] "NCBI data base," http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?accn =gse37815, series GSE37815.

[18] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," Pattern Recognition-Elsevier, vol. 30, no. 7, pp. 1145–1159, 1997.

[19] K. H. Zou, A. J. O. Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," Circulation, vol. 115, no. 5, pp. 654–657, February 2007.

[20] J. Hallinan, "Assessing and comparing classifier performance with roc curves," Macine Learning Mastery, 2014.

[21] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," Nucleic Acids Research, vol. 28, pp. 27–30, 2000.