DE GRUYTER
OPEN

JAISCR, 2013, Vol. 3, No. 2, pp. 101 – 115
DOI 10.2478/jaiscr-2014-0008

INTEGRATED STATISTICAL AND RULE-MINING TECHNIQUES FOR DNA METHYLATION AND GENE EXPRESSION DATA ANALYSIS

Saurav Mallik[1], Anirban Mukhopadhyay[2] and Ujjwal Maulik[3]

[1]*Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India*
E-mail: chasaurav_r@isical.ac.in, sauravmtech2@gmail.com

[2]*Department of Computer Science and Engineering, University of Kalyani, Kalyani, India*
E-mail: anirban@klyuniv.ac.in

[3]*Department of Computer Science and Engineering, Jadavpur University, Kolkata, India*
E-mail: umaulik@cse.jdvu.ac.in

Abstract

For determination of the relationships among significant gene markers, statistical analysis and association rule mining are considered as very useful protocols. The first protocol identifies the significant differentially expressed/methylated gene markers, whereas the second one produces the interesting relationships among them across different types of samples or conditions. In this article, statistical tests and association rule mining based approaches have been used on gene expression and DNA methylation datasets for the prediction of different classes of samples (viz., Uterine Leiomyoma/class-formersmoker and uterine myometrium/class-neversmoker). A novel rule-based classifier is proposed for this purpose. Depending on sixteen different rule-interestingness measures, we have utilized a Genetic Algorithm based rank aggregation technique on the association rules which are generated from the training set of data by Apriori association rule mining algorithm. After determining the ranks of the rules, we have conducted a majority voting technique on each test point to estimate its class-label through weighted-sum method. We have run this classifier on the combined dataset using 4-fold cross-validations, and thereafter a comparative performance analysis has been made with other popular rule-based classifiers. Finally, the status of some important gene markers has been identified through the frequency analysis in the evolved rules for the two class-labels individually to formulate the interesting associations among them.

1 Introduction

In the present scenario, it has been observed that certain epigenetic factors have a major role in the regulation of a gene. It has been proved that DNA methylation is an important epigenetic factor which can seize the transcription of a gene. Therefore, gene-expression levels are automatically biased to non-differential expression. The DNA methylation is observed in the promoter region of a gene. The methylated DNA is bound by methyl-CpG-binding domain proteins (MBDs) which might push some other proteins to the locus (viz., histone deacetylases and other chromatin remodeling proteins) that may alter histones. Hence, inactive chromatin (i.e., heterochromatin) has been formed and transcriptional silencing of hyper-methylated genes has been took place.

In this article, DNA methylation and gene expression [3], [4], [33] data have been analyzed through an integrated approach [5] of some statistical testing [7], [8], [21] and association rule mining [10], [12] based techniques. First a normality test has been applied on the data to know whether it is normally distributed. If so, then some parametric statistical tests have been utilized on it, otherwise some non-parametric tests are used to obtain differentially expressed (i.e., DE) [18], [19] or differentially methylated (i.e., DM) genes correctly. Thereafter, the intersected DE/DM genes are identified from all parametric tests as well as all non-parametric tests [14], [22] individually; and top common DE/DM genes are then chosen for the next step. The data discretization of the genes has been performed using k-means clustering sample-wise. The discretized data are then subdivided into test data and training data using 4-fold cross-validations (CVs). We have also applied updated Apriori rule-mining algorithm [11] on the training data and determined frequent closed itemsets (FCIs) at a minimum support value. From the itemsets, corresponding association rules have been extracted and evaluated w.r.t. 16 different rule-interestingness measures [17], [20]. On the basis of the 16 interestingness measures, we have performed a Genetic Algorithm (GA) based rank aggregation technique [2] on the evolved rules to estimate final ranking list of the evolved rules. A majority voting technique is then conducted on each test point to determine its class-label (i.e., either experimental/treated class-label or control/normal class-label) whose training rules maximally satisfy that test point through weighted-sum technique. Our proposed classifier is verified on one methylation and two expression datasets using 4-fold CVs. A comparative performance analysis has been done between our proposed classifier and the other popular rule-based classifiers.

Some significant observations have been finally demonstrated on the status of some important genes through the frequency analysis in association rules for the two class labels individually.

The remaining sections of the article are arranged as follows. Section 2 contains a description of our proposed methodology. The source and information of some real datasets have been enlisted in section 3. Section 4 reports the experimental results with discussion. Finally, section 5 draws the concluding remarks.

2 Proposed Methodology

In this section, we have described the proposed methodology in detail.

2.1 Normalization

Suppose, $MD[i, j]$ is input data matrix of the top genes, where i and j refer to genes and samples, respectively. First of all, the data should be normalized as normalization converts the data from different scales into a common scale. There are many normalization methods available. In our experiment, we have used zero-mean normalization [32] which converts the data into such configuration where mean of each row (i.e., each gene) becomes zero and standard deviation becomes one.

2.2 Normality Test

It is fact that parametric statistical tests perform well for the normally distributed data, and non-parametric statistical tests fit well for the data that do not follow normal distribution. Thus, normality test for each gene of each dataset is essential here. Therefore, Jarque-Bera test [25] has been applied on each gene as normality test.

2.3 Determination of DE/DM genes using different statistical tests

The parametric and non-parametric tests are then applied based on results of the normality test to identify DE/DM genes. In our experiment, we have used two parametric tests (viz., t-test and modified Bayes t-test [7]) and two non-parametric tests (viz., Limma [14], SAM). Thereafter, the intersection of all the parametric tests as well as non-parametric tests are obtained separately. The common genes are then ranked using majority voting technique based on their p-values for the different tests and top 160 genes are listed in total, in which 80 among them are upregulated/hypomethylated and rest of them are downregulated/hypomethylated.

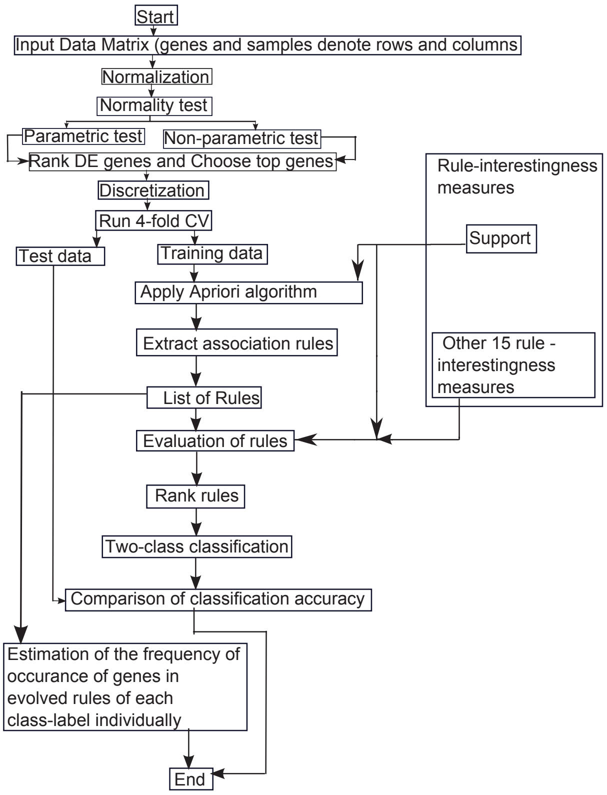

Start
Input Data Matrix (genes and samples denote rows and columns)
Normalization
Normality test
Parametric test
Non-parametric test
Rank DE genes and Choose top genes
Discretization
Run 4-fold CV
Test data
Training data
Apply Apriori algorithm
Extract association rules
List of Rules
Evaluation of rules
Rank rules
Two-class classification
Comparison of classification accuracy
Estimation of the frequency of occurance of genes in evolved rules of each class-label individually
End
Rule-interestingness measures
Support
Other 15 rule - interestingness measures

Figure 1. Flowchart of the whole methodology of the proposed rule-base classifier.

2.4 Data Discretization

The top genes are then discretized into boolean form. At first, we transpose *MD*. Let, *MDT* be the resulting matrix. At this moment, discretization of the input data matrix is mandatory for ARM. Thus, we run k-means clustering algorithm row-wise (i.e., sample-wise) on each row of *MDT* where number of clusters is set to 2. The cluster that has higher centroid value is considered to be cluster of up-regulated/hyper-methylated genes and other becomes cluster of down-regulated/hypo-methylated genes.

Number of columns of *MDT* is set to twice the original number of columns, where the first half of columns are for up-regulation/hyper-methylation property and the second half of columns are for down-regulation/hypo-methylation property. For the first half, 1 denotes upregulated/hypermethylated and 0 denotes downregulated/hypomethylated. Similarly, for the second half, 1 denotes downregulated/hypomethylated and 0 denotes upregulated/hypermethylated. Two extra columns are then added at the end of all columns, where first extra column refers to the treated class-label (i.e., class=tumor for Dataset 1 or class-formersmoker for Dataset 2) and second extra column refers to the normal/control class-label (i.e., class=normal for Dataset 1 or class-neversmoker for Dataset 2). The values of the experimental class-label are 1 for experimental samples and 0 for control samples. Similarly, the values of the control class-label are 1 for control samples and 0 for experimental samples. Here, we have shown an example of discretization of the input data matrix *MDT* in Table 1. Let us assume that *MDTb* is the resulting boolean matrix, whose rows refer to samples and columns denote genes. According to the table, s_{exp} and s_{ctr} denote experimental/treated and control samples respectively, where '+' and '-' denote up and down-regulation respectively.

Table 1. An example of discretization of data matrix: here, s_{exp} and s_{ctr} refer to experimental and normal samples respectively, '*Gn*' denotes gene, '+' and '-' denote hyper-methylation/up-regulation and hypo-methylation/down-regulation, respectively. 'Up-regulated region' and 'Down-regulated region' denote hyper-methylated/up-regulated and hypo-methylated/down-regulated regions, respectively.

	Up-regulated region			**Down-regulated region**			**Experimental class-label (Ex)**	**Control class-label (Ct)**
	Gn1+	**Gn2+**	**...**	**Gn1-**	**Gn2-**	**...**		
s_{exp1}	1	0	...	0	1	...	1	0
s_{exp2}	0	1	...	1	0	...	1	0
...	...	...	...	...	...	...	...	...
s_{ctr1}	1	1	...	0	0	...	0	1
s_{ctr2}	1	0	...	0	1	...	0	1
...	...	...	...	...	...	...	...	...

Here, for the first row/sample (i.e., s_{exp1}), gene1 (denoted by 'Gn1') is upregulated and gene2 (denoted by 'Gn2') is downregulated. Therefore, in the table, the cell $MDTb(s_{exp1}, Gn1+)$=1 as 'Gn1' is upregulated, and the cell $MDTb(s_{exp1}, Gn2-)$=1 as 'Gn2' is downregulated. Hence, automatically, the cells $MDTb(s_{exp1}, Gn1-)$=0 as 'Gn1' is not downregulated, and $MDTb(s_{exp1}, Gn2+)$=0 as 'Gn2' is not upregulated. Subsequently, $MDTb(s_{exp1}, Ex)$=1 as this row refers to an experimental/treated sample s_{exp1}. Similarly, $MDTb(s_{exp1}, Ct)$=0 as this row does not denote a control/normal sample. The similar technique is applicable for the other samples/rows of the matrix *MDTb*. In case of DM genes, the approach is same for methylation data of the table.

2.5 Cross validation

Subsequently, we divide *MDTb* data matrix into 4-folds, where one-fold of *MDTb* can be considered as test data, and remaining 3-fold data can be used as training data. This procedure will be repeated for 3 times again as it is 4-fold cross-validations.

2.6 A Novel ARM based approach

Before progressing further, let us discuss some fundamental concepts of ARM and Apriori algorithm [11]. ARM is a popular technique to estimate interesting relationships among different attributes (i.e., genes). It produces different association rules depending on most frequent attributes. Suppose, $It = \{i_1, i_2, ..., i_n\}$ be an itemset and $Trc = \{t_1, t_2, ..., t_m\}$ be a set of transactions. Thus, a rule might be described as $An \Rightarrow Cn$, where $An, Cn \subseteq It$ and $An \bigcap Cn = \phi$. Here, An is called as antecedent (i.e., set of items in left-hand side of a rule) and Cn is called as consequent (i.e., set of items in right-hand side of a rule). The support of It can be defined as the total number of transactions in which all items of It appear. It is frequent when its support is greater than a threshold value (i.e., minimum support). The confidence of a rule is defined as the ratio of support of $An \cap Cn$ to the support of An. Apriori is a basic algorithm for learning association rules. Apriori utilizes a "bottom-up" approach, where frequent subsets are extended one item at a time to generate each candidate and groups of the candidates are tested against the data. The algorithm terminates if there is no further successful extensions to be identified. The output of Apriori is actually the sets of rules that generate the occurrence of items in the dataset. Apriori follows a breadth-first search and a Hash tree structure to count the candidate itemsets. Apriori produces candidate itemsets of length k from itemsets of length $k-1$. Thereafter, it eliminates the candidates having an infrequent sub-pattern. The candidate set contains all frequent-itemsets. It searches the transaction database to discover most frequent itemsets from the candidates.

In our experiment, updated Apriori rule-mining algorithm [31] has been applied on the training subpart of *MDTb* and estimate FCIs with atleast two genes/items at 0.1 minimum support. The corresponding association rules have been extracted and evaluated with sixteen different rule-interestingness measures (i.e., support, confidence, hyperconfidence, lift, oddsRatio, leverage, conviction, cosine, doc, fishersExactTest, coverage, gini, improvement, phi, RLD and hyperLift) from CFIs. Each resulting rule must be of a special type (i.e., classification rule type), where consequent of it consists of class-label (viz., class=tumor/class-formersmoker, or class=normal/class-neversmoker) only. For gene expression data, the rules are like the followings (from the first two rows of Table 1):
$[Gn1+, Gn2-, ...] \Rightarrow [class = tumor]$
$[Gn1-, Gn2+, ...] \Rightarrow [class = normal]$,
where 'Gn' denotes gene, '+' denotes up-regulation and '-' denotes down-regulation. The first rule can be interpreted as follows: if gene1 is up-regulated and gene2 is down-regulated simultaneously, the condition or corresponding class label becomes experimental. The second rule may be defined as follows: if gene1 is down-regulated and gene2 is up-regulated simultaneously, the condition will be normal/control. Similarly, for DNA methylation data, types of rules will be same as expression data except '+' and '-' denote hyper and hypo-methylation respectively, instead of up and down-regulation. The rules are then ranked according to each of the rule-interestingness measures separately.

GA based rank aggregation algorithm [2] is then applied on the resulting rankings of the rules depending on the rule-interestingness measures, and the final ranking list of the rules has been obtained.

2.7 Two-class classification technique

Thereafter, two-class classification technique is needed to apply on the test data points. Thus, we have applied a majority voting technique on each test data point to determine its class label through weighted-sum method. Hence, we have assigned some weight (i.e., $0.0000 < weight < 1.0000$) in descending order on the final list of rules in such a way that the topmost rule gets the highest weight, 2nd topper gets 2nd highest weight and so on. Finally, for the lowest ranked rule, the lowest weight is assigned. The weight-interval between any two consecutive ranked rules is kept same here. Now, consider one test data point. We have filtered the association rules with their class-labels and corresponding weights by which the test data point is completely satisfied. Thereafter, we have added the weights of the rules whose original class-label is experimental and also do the same for the rules whose original class-label is control. The two results are then compared and the class-label with higher weighted-sum becomes the predicted class-label of the test data point. But, if the weighted-sum for one class-label is equal to the other, the class-label of the top rule that satisfies the test data point, becomes the predicted class-label of it. In case, if there is no such rule that satisfies the test data point, consider the class-label of the rule (belongs to the all rules of the dataset) that satisfies maximum number of test points, is considered as the predicted class-label of it.

After that, a performance analysis has been conducted. We calculate the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), and sensitivity [24], specificity [24], accuracy [24] and Mathews correlation coefficient (MCC) [24]. Sensitivity, specificity, accuracy and MCC are defined in the Equations 1, 2, 3 and 4, respectively.

$$sensitivity = \frac{TP}{(TP+FN)}, \tag{1}$$

$$specificity = \frac{TN}{(FP+TN)}, \tag{2}$$

$$accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}, \tag{3}$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{4}$$

2.8 Comparative performance analysis between proposed and other rule based classifiers

The performance of the proposed technique has been compared with that of some existing rule-based classifiers. Flowchart of the complete methodology of the proposed rule-base classifier is depicted in Figure 1.

3 Real Datasets

Here, two real datasets have been utilized which are described in Table 2. For the first dataset, we have used the common set of genes which have both expression values and methylation values.

4 Experimental Results and Discussion

4.1 Identification of DE/DM genes

For Dataset 1, we have considered only common genes (i.e, 13072 genes) that have both expression and methylation data, and then utilized our proposed method on the common genes for expression as well as methylation data individually. First of all, the datasets are normalized. Fig. 2(a) shows the data representation before using any normalization, and Fig. 2(b) depicts the data representation after using zero-mean normalization. 10,236 and 2,836 genes are found to be normal and non-normally distributed respectively for the expression data of Dataset 1, where for methylation data of Dataset 1, these numbers are 8,173 and 4,899 respectively. For Dataset 2, these are 15,113 and 7,170 respectively. Fig. 3(a) presents a normality plot for a normally distributed data of a gene, where Fig. 3(b) denotes the normality plot for a non-normally distributed data of another gene from Dataset 1.

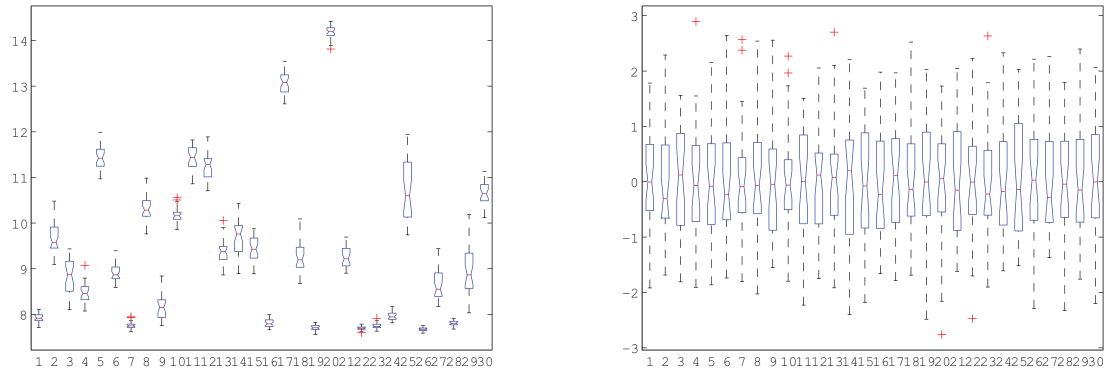
After normality test, different parametric and non-parametric tests are used. In Table 3, we have listed number of DE/DM genes by the statistical tests (at 0.05 p-value cutoff) and common DE/DM genes from the tests for each of the datasets. A volcanoplot by Limma test to identify DE genes and clustergram for expression data for Dataset 1 are depicted in Fig 4(a) and (b), respectively.

According to Table 3, we have filtered the common genes from each dataset, and ranked these using majority voting technique. Thereby, top forty genes from normally distributed up-regulated/hyper-methylated list (e.g., top 40 from 323 common DE genes in Table 3(a)) and another top forty from non-normally distributed up-regulated/hyper-methylated list (e.g., top 40 from 86 common DE genes in Table 3(b)) are selected. Similar thing is done for the down-regulated/hypo-methylated genes.

Hence, the top 160 DE genes (i.e., 40 plus 40 from up-regulated list; and 40 plus 40 for down-regulated list for each dataset) are finally chosen for the data discretization.

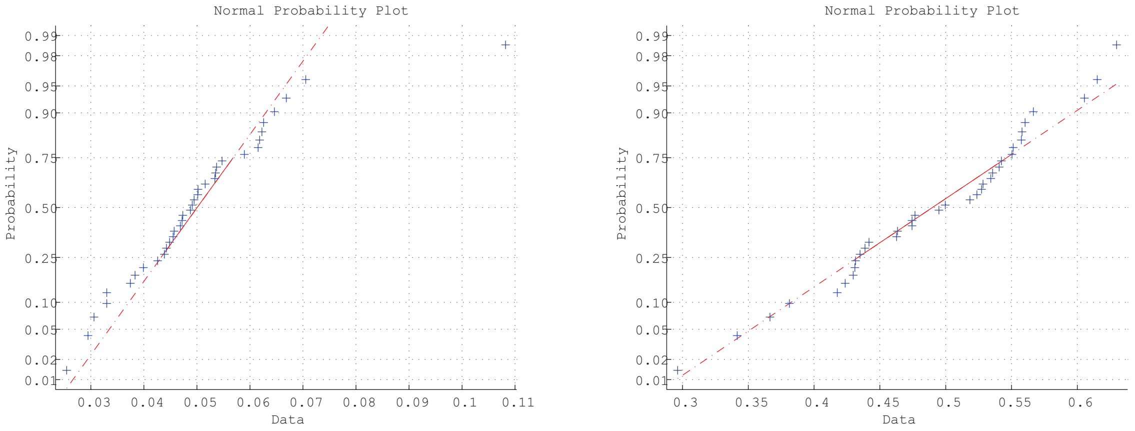
Table 2. Information of used Datasets (DS).

DS id	**Dataset information**	**NCBI Ref. id**	**Sample size**
1	Gene expression and genome-wide methylation datasets of Uterine Leiomyoma [1], having Uterine Leiomyoma tumor (experimental) and normal myometrial samples.	**GSE31699**	**32 (having 16 experimental and 16 normal).**
2	Gene expression dataset of cigarette smokers of lung adenocarcinoma [6], former smoker (FS) and never smoker (NS) for Non-Tumor samples.	**GSE10072**	**33 (having 18 FS and 15 NS).**



(a) (b)

Figure 2. Boxplots of data for each gene of Dataset 1 (a) before, and (b) after zero-mean normalization, where each boxplot denotes data for each gene.(Here, boxplots for the first thirty genes are presented.)



(a) (b)

Figure 3. Normality plots for (a) normally distributed expression data of a gene, and (b) non-normally distributed expression data for another gene, of Dataset 1.

Table 3. Number of differentially expressed/methylated genes by different statistical tests (at 0.05 p-value cutoff)- for the expression data of (a) normally distributed genes, (b) non-normally distributed genes of Dataset 1; -for the methylation data of (c) normally distributed genes, (d) non-normally distributed genes of Dataset 1; and the expression data (e) normally distributed genes, (f) non-normally distributed genes of Dataset 2. Here $\#gene_{up}$, $\#gene_{down}$ denote up and down-regulated genes respectively; where $\#gene_{hyper}$ and $\#gene_{hypo}$ refer to hyper and hypo-methylated genes respectively.

(a)

	t-test	**Bayes t-test**	**common**
$\#gene_{up}$	391	329	323
$\#gene_{down}$	576	491	486

(b)

	SAM	**Limma**	**common**
$\#gene_{up}$	86	86	86
$\#gene_{down}$	70	70	70

(c)

	t-test	**Bayes**	**common**
$\#gene_{hyper}$	695	548	548
$\#gene_{hypo}$	686	608	608

(d)

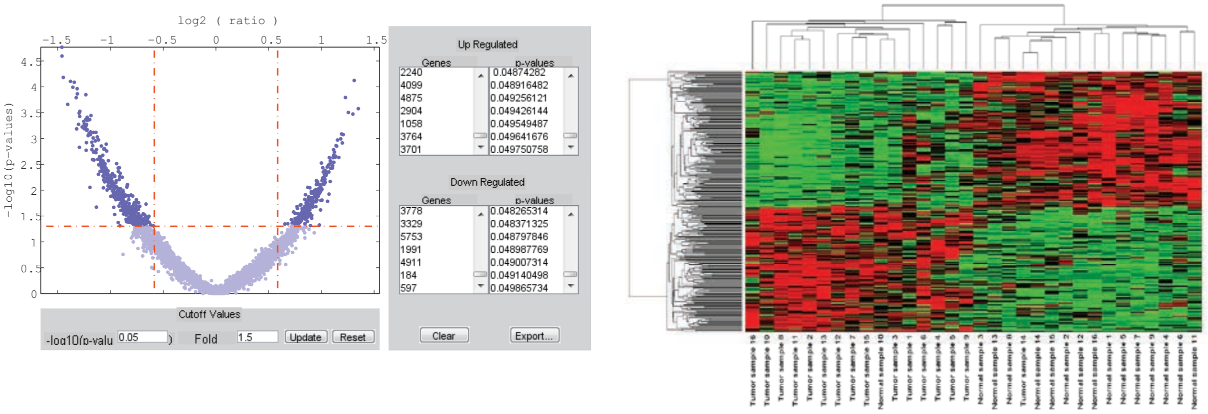
	SAM	**Limma**	**common**
$\#gene_{hyper}$	174	177	174
$\#gene_{hypo}$	165	170	165

(e)

	t-test	**Bayes t-test**	**common**
$\#gene_{up}$	462	88	46
$\#gene_{down}$	536	93	46

(f)

	SAM	**Limma**	**common**
$\#gene_{up}$	200	201	185
$\#gene_{down}$	290	238	233



Up Regulated

Genes	p-values
2240	0.04874282
4099	0.048916482
4875	0.049256121
2904	0.049426144
1058	0.049549487
3764	0.049641676
3701	0.049750758

Down Regulated

Genes	p-values
3778	0.048265314
3329	0.048371325
5753	0.048797846
1991	0.048987769
4911	0.049007314
184	0.049140498
597	0.049865734

(a) (b)

Figure 4. (a) Volcanoplot by Limma test for identifying differentially expressed genes and (b) clustergram, for gene expression data of Dataset 1.

Table 4. Comparative performance analysis of the rule based classifiers on (a) gene expression data, and (b) methylation data for Dataset 1.

	(a)				(b)			
Rule based classifier	**Average sensitivity [%](s.d.)**	**Average specificity [%](s.d.)**	**Average accuracy [%](s.d.)**	**Average MCC [%](s.d.)**	**Average sensitivity [%](s.d.)**	**Average specificity [%](s.d.)**	**Average accuracy [%](s.d.)**	**Average MCC [%](s.d.)**
Proposed	72.92 (7.22)	**89.58** (6.99)	**81.25** (6.25)	**0.64** (0.13)	91.67 (3.61)	**100** (0.00)	**95.83** (1.80)	**0.92** (0.03)
Conjunctive-Rule	**89.58** (0.04)	62.5 (0.06)	76.04 (4.77)	0.54 (0.09)	**100** (0.00)	68.75 (0.06)	84.38 (3.13)	0.72 (0.09)
Decision-Table	87.5 (0.00)	62.5 (0.11)	75 (5.41)	0.52 (0.10)	**100** (0.00)	81.25 (0.00)	90.63 (0.00)	0.83 (0.00)
JRip	83.33 (0.07)	64.58 (0.10)	73.96 (6.50)	0.49 (0.13)	**100** (0.00)	91.67 (0.07)	95.83 (3.61)	0.92 (0.07)
OneR	87.5 (0.00)	64.58 (0.10)	76.04 (4.77)	0.54 (0.09)	**100** (0.00)	81.25 (0.00)	90.63 (3.61)	0.83 (0.00)
PART	83.33 (0.07)	68.75 (0.11)	76.04 (7.86)	0.53 (0.16)	**100** (0.00)	91.67 (0.07)	95.83 (3.61)	0.92 (0.07)
Ridor	77.08 (0.04)	77.08 (0.13)	77.08 (7.22)	0.54 (0.14)	**100** (0.00)	91.67 (0.07)	92.71 (1.80)	0.86 (0.03)

4.2 Comparative performance analysis between proposed and other rule based classifiers

For comparison purpose, we have taken into account other popular rule-based classifiers (i.e., ConjunctiveRule, DecisionTable, JRip, OneR, PART, Ridor) implemented in Weka 3.6 software and then we have compared the performance of our proposed classifier with the other rule based classifiers. In our experiment, we have run 4-fold CV for 3 times and then we have estimated average sensitivity, average specificity, average accuracy and average MCC. Tables 4 (a) and (b) report the comparative performance among the classifiers for the expression and methylation data respectively for Dataset 1. Similarly, Table 5 presents the same for Dataset 2.

For the expression data of Dataset 1, our proposed classifier provides the better average accuracy (i.e., 81.25%) and better average MCC (i.e., 0.64) than the other rule based classifiers (see Table 4(a)). For methylation data of the Dataset, our classifier provides better average accuracy (i.e., 95.83%) and average MCC (i.e., 0.92) than ConjunctiveRule, DecisionTable, OneR and Ridor classifiers. JRip and PART yield same average accuracy and average MCC as the proposed one (see Table 4(b)). But, from the point of standard deviation (s.d.) of accuracies, proposed one is better than the two classifiers as the s.d. of accuracies (i.e., 1.80) of our classifier is lower than the other two. Similarly, w.r.t. s.d. of MCCs, it is also better than the other two as the s.d. of MCCs of it (i.e., 0.03) is lower than the others. For above expression and methylation data of Dataset 1, average specificity of the proposed one is better than the others, but sensitivity of it is lower than the others. In case of Dataset 2, our classifier also performs better than others in term of average specificity, average accuracy and average MCC (see Table 5). The sensitivity of it is found less than Ridor classifier, but better than others. Moreover, Fig. 5 presents barplot for the datasets comparing the average accuracies of the classifiers.

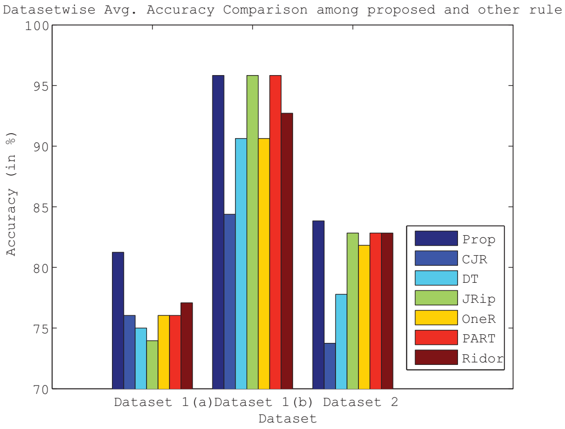


Figure 5. Comparison of accuracies among different statistical tests for (a) expression and (b) methylation data of Dataset 1, and (c) expression data of Dataset 2, respectively, where ‘Prop’, ‘CJR’ and ‘DT’ denote proposed, ConjunctiveRule and DecisionTable classifiers respectively.

In fact, from expression data, we have got total 744 (i.e., 428 plus 316) association rules for all 4-fold CVs, where the rules having ‘class=tumor’ in the consequent are found 428 times and the rules having ‘class=normal’ in the consequent are found 316 times.

For the expression data of Dataset 1, we have observed that down-regulation of ‘TACSTD2’ gene is found with 99.77% frequency of occurring in the evolved rules of tumor class-label. But, down-regulation of the gene has not been found in the evolved rules of control class-label. Therefore, down-regulation of it seems to be extremely important for tumor (i.e., Uterine Leiomyoma) formation. We have found some literature-base evidences about ‘TACSTD2’ gene for tumor formation in [23], [30]. Down-regulation of another gene ‘ACSL5’ has been identified with 73.36% frequency of occurring in the evolved rules of tumor class-label; but down-regulation of it has not been found in the evolved rules of control class-label. Thereby, down-regulation of ‘ACSL5’ is also important like down-regulation of ‘TACSTD2’ for Uterine Leiomyoma formation. Down-regulation of ‘FHL5’ has some less importance in tumor formation as down-regulation of it is found with .63% fre-

Table 5. Comparative performance analysis of the rule based classifiers on gene expression data od Dataset 2.

Rule based classifier	Average sensitivity [%](s.d.)	Average specificity [%](s.d.)	Average accuracy [%](s.d.)	Average MCC [%](s.d.)
Proposed	85.96 (6.07)	**80.95** (4.12)	**83.84** (4.63)	**0.67** (0.09)
Conjunctive-Rule	73.85 (4.32)	76.82 (3.03)	73.74 (4.63)	0.50 (0.07)
Decision-Table	79.67 (9.46)	76.82 (3.03)	77.78 (7.63)	0.57 (0.13)
JRip	84.21 (5.26)	**80.95** (4.12)	82.83 (4.63)	0.65 (0.09)
OneR	84.21 (5.26)	78.57 (0)	81.82 (3.03)	0.63 (0.06)
PART	87.31 (2.75)	77.38 (2.06)	82.83 (1.75)	0.65 (0.03)
Ridor	**89** (4.58)	76.45 (3.67)	82.83 (1.75)	0.66 (0.03)

Table 6. Occurrence of highly-frequent genes in evolved rules of experimental and normal class-labels for 4-fold CVs in (a) gene expression data and (b) methylation data, both from Dataset 1 and (c) gene expression data from Dataset 2, respectively; here '+' denotes up-regulation/hyper-methylation and '-' denotes down-regulation/hypo-methylation.

(a)

Gene	Frequency [%] of occurrence in evolved rules of 'class=tumor'	Frequency [%] of occurrence in evolved rules of 'class=normal'
TACSTD2-	99.77	0
ACSL5-	73.36	0
FHL5-	100	0.63
CALCRL+	0.23	93.04

(b)

Gene	Frequency [%] of occurrence in evolved rules of 'class=tumor'	Frequency [%] of occurrence in evolved rules of 'class=normal'
NRTN-	100	0
PRSS8+	100	14.19

(c)

Gene	Frequency [%] of occurrence in evolved rules of 'class-formersmoker'	Frequency [%] of occurrence in evolved rules of 'class-neversmoker'
204224_s_at-	96.23	0
207968_s_at+	94.21	0
209780_at+	0	95.67
209717_at+	0	62.68
221578_at-	0	61.93

Table 7. The p-values of top frequent genes for (a) expression data, (b) methylation data, both for Dataset 1 and (c) expression data, for Dataset 2, respectively.

(a)

Gene	**p-value**
TACSTD2 **(Non-normally distributed gene)**	1.46E-05 (in Limma), 0.000353 (in SAM).
ACSL5 **(Non-normally distributed gene)**	2.50E-05 (in Limma), 0.000353 (in SAM).

(b)

Gene	**p-value**
NRTN **(Normally distributed gene)**	1.66E-12 (in t-test), 5.85E-14 (in Bayes t).

(c)

Gene	**p-value**
204224_s_at **(Non-normally distributed gene)**	0.0148 (in Limma), 0.0068 (in SAM).
207968_s_at **(Non-normally distributed gene)**	0.0346 (in Limma), 0.0326 (in SAM).
209780_at **(Non-normally distributed gene)**	0.02565 (in Limma), 0.02185 (in SAM).
209717_at **(Non-normally distributed gene)**	0.0069 (in Limma), 0.0052 (in SAM).
221578_at **(Non-normally distributed gene)**	0.0419 (in Limma), 0.0366 (in SAM).

quency of occurring in the evolved rules of control class-label in spite of 100% frequency of occurring in the evolved rules of tumor class-label. Similarly, up-regulation of 'CALCRL' has some less importance for non-tumorous condition as up-regulation of the gene is already identified with .23% frequency of occurring in the evolved rules for tumor class-label in spite of 93.04% such frequency for control class-label.

For the methylation data of Dataset 1, we have found total 399 (i.e., 251 plus 148) association rules, where the rules having 'class=tumor' in the consequent are identified 251 times and the rules having 'class=normal' in the consequent are identified 148 times. Here, we have found that hypo-methylation of 'NRTN' is identified with 100% frequency of occurring in the evolved rules of tumor class-label for both 4-fold CVs. Hypo-methylation of the gene is not found in the evolved rules of control class-label. So, hypo-methylation of it seems to be very important for Uterine Leiomyoma formation. Hyper-methylation of 'PRSS8' is also found with 100% frequency for tumor class-label and 14.19% frequency for control class-label. Therefore, it is less responsible for tumor formation.

Similarly, for the expression data of Dataset 2, we have obtained 96.23% frequency of down-regulated gene 204224_s_at- (here, it is probe-id of a gene) for class-formersmoker, but nothing for class-neversmoker. Hence, down-regulation of 204224_s_at- is important for the class-formersmoker. Similarly, up-regulated gene 207968_s_at+ is also important for the class-formersmoker. Subsequently, 209780_at+, 209717_at+ and 221578_at- are also significant for class-neversmoker. For details, see Table 6(a), (b) and (c) respectively for the datasets. The p-values of the top significant genes in different statistical tests are listed in Table 7(a), (b) and (c), respectively for the datasets.

Finally, we have applied GA based rank aggregation algorithm on all evolved rules depending upon the rule-interestingness measures, and top 10 rules from the results are listed in Table 9(a), (b) and (c) respectively for all the datasets.

4.3 Integrated analysis of the expression and methylation data for Dataset 1

Here we discuss the integrated analysis on the gene expression and methylation data for Dataset 1. We have already found the 409 up-regulated genes (i.e., 323 from normally distributed data of genes plus 86 from non-normally distributed data of genes), and 556 down-regulated genes (i.e., 486 from normally distributed data of genes plus 70 from non-normally distributed data of genes) from the expression data. Subsequently, 692 hyper-methylated genes (i.e., 507 from normally distributed data of genes plus 185 from non-normally distributed data of genes), and 765 hypo-methylated genes (i.e., 600 from normally distributed data of genes plus 165 from non-normally distributed data of genes) have been determined from the methylation data. Suppose, number of such up and down-regulated, hyper and hypo-methylated genes are denoted as $\#G_{up}$, $\#G_{down}$, $\#G_{hyper}$ and $\#G_{hypo}$, respectively. Thereby, 17 genes have been identified as $\#(G_{up} \cap G_{hyper})$, where 34 genes as $\#(G_{up} \cap G_{hypo})$ (see Fig. 6(a) and Table 8). Similarly, 62 and 22 are found as $\#(G_{down} \cap G_{hyper})$ and $\#(G_{down} \cap G_{hypo})$, respectively (see Fig. 6(b) and Table 8).

Hence, we have identified two types of inverse co-relationships between the expression data and methylation data for the above cases, one is for $\#(G_{up} \cap G_{hypo})$ where 34 genes are identified, and other is for $\#(G_{down} \cap G_{hyper})$ where 62 are detected.

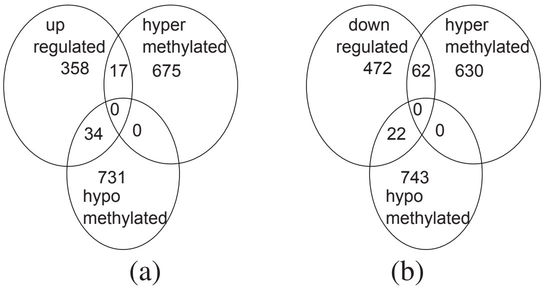


Figure 6. Intersections of any two among: (a) $\#G_{up}$, $\#G_{hyper}$ and $\#G_{hypo}$, and (b) $\#G_{down}$, $\#G_{hyper}$ and $\#G_{hypo}$, respectively for Dataset 1.

Therefore, according to these observations, we draw conclusions that the 34 genes are up-regulated due to the hypo-methylation effect on the genes. Similarly, the 62 genes are down-regulated due to heavy methylation effect on the genes. But, in case of the 17 genes, there might be some other epigenetic effects which dominate the hyper-

Table 8. (a) $G_{up} \cap G_{hyper}$, (b) $G_{up} \cap G_{hypo}$, (c) $G_{down} \cap G_{hyper}$, (d) $G_{down} \cap G_{hypo}$ respectively for Dataset 1.

	#Genes	**Genes**
4* $G_{up} \cap G_{hyper}$	17	APBA2, C1orf61, CNKSR1, DBC1, DSG2, ETNK2, FKBP7, HDAC3, LHFPL2, NOPE, PCSK1, PNMA3, PRKD1, THEG, UAP1L1, UNC5D, ZIM2.
8* $G_{up} \cap G_{hypo}$	34	B4GALNT4, BCAN, BSN, C20orf100, CAD, CDO1, COL6A3, DDB2, FSD1, GALNT13, GAP43, GDF15, GLIS1, GPT2, H2AFY, HOXA11, HSD17B6, KCNG1, KLHL13, MAMDC4, PDE8B, PHF13, PHLDB2, PI15, RPE65, SCN2B, SEMA7A, SHOX2, TDO2, TH, THSD4, TNFSF4, TP53INP1, TUBB3.
8* $G_{down} \cap G_{hyper}$	62	ABI3, BMX, BST2, C11orf52, C1orf115, C8orf4, C9orf58, CALCRL, CARD10, CCDC68, CD2, CD34, CD40, CD52, CD79B, CD8A, CLDN5, CLIC1, CMTM8, CREG1, CRIM1, CRIP1, CYBA, CYTL1, EDG1, EGFL7, EMCN, EVI1, FBLN2, GIMAP5, GRAMD3, HLA-DMA, HOXB8, ICAM2, ICAM3, IFI27, ITGB7, KLF11, LRP5, LYST, MFNG, MFSD7, MMRN2, MTSS1, MVP, MYOT, NUAK1, PCDHGC4, PECAM1, RASIP1, RPH3AL, SCN4B, SH2D3C, SLC25A18, SLC35A3, SQRDL, ST6GALNAC1, STEAP4, TEK, TMC6, TMEM71, ZNF217.
6* $G_{down} \cap G_{hypo}$	22	ACSL5, ARHGAP9, CYB5R3, GDPD5, GFOD1, HSPB2, IL7R, LRRC51, MAPK10, NFS1, OR51E2, PAM, PKHD1, PTPRCAP, RAMP3, RHAG, SDC4, SEMA3B, SLAMF6, SORBS2, STARD8, TESC.

methylation effect totally on the genes. That's why, the genes are still up-regulated. Similarly, as the hypo-methylation effect is completely dominated by other epigenetic effects on the 22 genes, the genes are still down-regulated in spite of the hypo-methylation effect.

5 Conclusion

An integrated analysis of statistical methodologies and ARM has been performed on gene expression and DNA methylation data for the prediction of experimental (i.e., Uterine Leiomyoma/class-formersmoker) and control (i.e., Uterine myometrium/class-neversmoker) class-labels. Some important observations on the combined dataset are also made applying our integrated analysis. Moreover, we have proposed a novel rule based classifier. Based on the sixteen different rule-interestingness measures, we have also applied GA based rank aggregation technique on the association rules that are generated from the training set of data by Apriori algorithm. After determining the ranks of the rules, we have conducted a majority voting technique on each test point to determine its class-label (i.e., experimental or control class-label) through weighted-sum method. We have run this classifier on the combined dataset using 4-fold CVs. Moreover, a comparative performance analysis is conducted between our proposed classifier and other existing rule based classifiers. Finally, we have predicted the status of some significant genes through the frequency analysis in the evolved rules for the two class-labels individually.

References

[1] A. Navarro, P. Yin, D. Monsivais, S. M. Lin, P. Du, J. J. Wei, S. E. Bulun, Genome-Wide DNA Methylation Indicates Silencing of Tumor Suppressor Genes in Uterine Leiomyoma, PLoS One, vol. 7, no. 3, pp. e33284, 2012.

[2] V. Pihur, S. Datta and S. Datta, RankAggreg, an R Package for Weighted Rank Aggregation, BMC Bioinformatics, vol. 10, pp. 62-72, 2009.

[3] K.R.V. Eijk, S.D. Jong, M.P.M. Boks, T. Langeveld, F. Colas, J.H. Veldink, C.G.F.D. Kovel, E. Janson, E. Strengman, P. Langfelder, R. S. Kahn, L. H. V. D. Berg, S. Horvath and R. A. Ophoff, Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects, BMC Genomics, vol. 13, no. 636, 2012.

Table 9. Top 10 association rules of genes of experimental and normal class-labels in (a) gene expression data, and (b) methylation data of Dataset 1; and (c) gene expression data of Dataset 2, respectively.

(a)

{BAX+, ISG20L1+, FHL5+, ACSL5+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}
{BAX+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC-, STEAP2-, HS3ST1- ⇒ class=normal}
{BAX+, ACSL5+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}
{BAX+, GALNT13-, NTNG1-, GLIS1-, HMGB3-, LIG1-, EDA2R-, FHL5-, OTC-, TACSTD2-, ACSL5-, STEAP2-, HRC-, HS3ST1- ⇒ class=tumor}
{BAX+, ISG20L1+, GALNT13-, GLIS1-, LIG1-, EDA2R-, FHL5-, OTC-, TACSTD2-, ACSL5-, STEAP2-, CALCRL-, HRC-, HS3ST1- ⇒ class=tumor}
{BAX+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC-, HS3ST1- ⇒ class=normal}
{BAX+, ACSL5+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}
{BAX+, ACSL5+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}
{BAX+, ACSL5+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}
{BAX+, CALCRL+, TNFSF10+, PRKCH+, GALNT13-, NTNG1-, GLIS1-, IGF2-, GDF15-, HMGB3-, LIG1-, EDA2R-, OTC- ⇒ class=normal}

(b)

{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, PMF1+, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, CCR1-, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, CCR1-, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, CCR1-, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, PANX3+, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, ENTPD3-, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{SEMA4G+, H19+, IL29+, TMEM71+, GP9+, NRTN+, CFHR1+, PMF1+, PANX3+, ENTPD3-, CCR1-, C9orf58-, LDB3-, FBLIM1-, TUBB3- ⇒ class=normal}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}
{PRSS8+, SEMA4G+, H19+, IL29+, TMEM71+, GP9+, CCDC13+, C9orf58-, NRTN-, CD1A-, LDB3-, KIAA1641-, NAV1-, FBLIM1-, TUBB3- ⇒ class=tumor}

(c)

{201449_at+, 204976_s_at+, 209717_at+, 218152_at+, 211698_at+, 206272_at+, 219025_at-, 218475_at- ⇒ class-neversmoker}
{205941_s_at+, 211071_s_at+, 201562_s_at+, 220688_s_at+ ⇒ class-formersmoker}
{210977_s_at+, 220213_at+, 205142_x_at+, 211181_x_at+, 201982_s_at+, 208916_at+, 207312_at+, 208096_s_at-, 204224_s_at- ⇒ class-formersmoker}
{38892_at+, 202251_at+, 201449_at+, 204976_s_at+, 209717_at+, 218152_at+, 211698_at+, 206272_at+, 219025_at-, 218475_at- ⇒ class-neversmoker}
{205941_s_at+, 215809_at+, 212016_s_at+, 211071_s_at+, 201562_s_at+, 220688_s_at+, 214198_s_at-, 204224_s_at-, 212281_s_at- ⇒ class-formersmoker}
{65588_at+, 205941_s_at+, 219256_s_at+, 211071_s_at+, 210405_x_at+, 205574_x_at+, 201562_s_at+, 220688_s_at+, class-formersmoker}
{210977_s_at+, 205142_x_at+ ⇒ class-formersmoker }
{214146_s_at+, 210237_at-, 202250_s_at- ⇒ class-neversmoker }
{211698_at+, 208744_x_at- ⇒ class-neversmoker}
{65588_at+, 202507_s_at+, 210977_s_at+, 215809_at+, 205142_x_at+, 217193_x_at+, 211380_s_at+, 204199_at+ ⇒ class-formersmoker}

[4] C.C. Yu, M. Furukawa, K. Kobayashi, C. Shikishima, P.C. Cha, J. Sese, H. Sugawara, K. Iwamoto, T. Kato, J. Ando and T. Toda, Genome-Wide DNA Methylation and Gene Expression Analyses of Monozygotic Twins Discordant for Intelligence Levels, PLoS One, vol. 7, no. 10, pp. e47081, 2012.

[5] S. Mallik, A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, Integrated Analysis of Gene Expression and Genome-wide DNA Methylation for Tumor Prediction: An Association Rule Mining-based Approach, Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE Symposium Series on Computational Intelligence - SSCI 2013, Singapore, pp. 120-127, April 16, 2013.

[6] M.T. Landi, T. Dracheva, M. Rounno, J.D. Figueroa, H. Liu, A. Dasgupta, F.E. Mann, J. Fukuoka, M. Hames, A.W. Bergen, S.E. Murphy, P. Yang, A.C. Pesatori, D. Consonni, P.A. Bertazzi, S. Wacholder, J.H. Shih, N.E. Caporaso and J. Jen, Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival, PLoS One, vol. 3, no. 2, pp. e1651, 2008.

[7] R. J. Fox, M. W. Dimmic, A Two-Sample Bayesian t-test for Microarray Data, BMC Bioinformatics, vol. 7, no. 126, pp. 1-11, 2006.

[8] A. Vickers, Parametric Versus Non-Parametric Statistics in the Analysis of Randomized Trials with Non-Normally Distributed Data, BMC Medical Research Methodology, vol. 5, no. 35, 2005.

[9] A. Mukhopadhyay, U. Maulik, and S. Bandyopdhyay, On Biclustering of Gene Expression Data, Current Bioinformatics, vol. 5, no. 3, pp. 204-216, 2010.

[10] A. Mukhopadhyay, U. Maulik, and S. Bandyopdhyay, A Novel Biclustering Approach to Association Rule Mining for Predicting HIV-1-Human Protein Interactions, PLoS One, vol. 7, no. 4, pp. e32289, 2012.

[11] R. Agrawal, T. Imielinski and A. Swami, Mining Association Rules between Sets of Items in large Databases, In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMOD'93), New York, NY, USA: ACM, pp. 207-216, 1993.

[12] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay and R. Eils, Mining Association Rules from HIV-human Protein Interactions, Int Conf. Systems in Medicine and Biology (ICSMB), pp. 344-348, 2010.

[13] W. H. Catherino, C. Prupas, J. C. Tsibris, P. C. Leppert and M. Payson, Strategy for Elucidating Differentially Expressed Genes in Leiomyomata Identified by Microarray Technology, Fertil Steril, vol. 80, pp. 282-290, 2003.

[14] G. Smyth, Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, Statistical Applications in Genetics and Molecular Biology, vol. 3, no. 1, pp. 3, 2004.

[15] P. Sethi and S. Alagiriswamy, Association Rule Based Similarity Measures for the Clustering of Gene Expression Data, The Open Medical Informatics Journal, vol. 4, pp. 63-73, 2010.

[16] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo and A. Pascual-Montano, Integrated Analysis of Gene Expression by Association Rules Discovery, BMC Bioinformatics, vol. 7, no. 54, 2006.

[17] X. Li, S. Mabu, H. Zhou, K. Shimada and K. Hirasawa, Analysis of Various Interestingness Measures in Class Association Rule Mining, SICE Journal of Control, Measurement, and System Integration, vol. 4, no. 4, pp. 295-304, 2011.

[18] C. Creighton, and S. Hanash, Mining Gene Expression Databases for Association Rules, Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.

[19] F. Tao, Weighted Association Rule Mining using Weighted Support and Significance Framework, In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C., USA, pp. 661-666, 2003.

[20] M. Anandhavalli, M. K. Ghose, and K. Gauthaman, Interestingness Measure for Mining Spatial Gene Expression Data using Association Rule, Journal of Computing, vol. 2, no. 1, pp. 110-114, 2010.

[21] S. Dudoit, Y. Yang, T. Speed, and M. Callow, Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments, Statistica Sinica, vol. 12, pp. 111-139, 2002.

[22] S. Y. Kim, J. W. Lee, and I. S. Sohn, Comparison of Various Statistical Methods for Identifying Differential Gene Expression in Replicated Microarray Data, Statistical Methods in Medical Research, vol. 15, pp. 3-20, 2006.

[23] X. Luo, L. Ding, J. Xu and N. Chegini, Gene Expression Profiling of Leiomyoma and Myometrial Smooth Muscle Cells in Response to Transforming Growth Factor-β, Endocrinology, vol. 146, no. 3, pp. 1097-1118, March 2005.

[24] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner, False Discovery Rate, Sensitivity and Sample Size for Microarray Studies, Bioinformatics, vol. 21, pp. 3017-3024, 2005.

[25] C.M. Jarque and A.K. Bera, A test for normality of observations and regression residuals, International Statistical Review, vol. 55, no. 2, pp. 163-172, 1987.

[26] Z. Wang and V. Palade, Building Interpretable Fuzzy Models for High Dimentional Data Analysis in Cancer Diagnosis, BMC Genomics, no. 12(S2):S5, 2011.

[27] Z. Wang, V. Palade and Y. Xu, Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis, In Proceedings of the 2006 International Symposium on Evolving Fuzzy Systems, IEEE 2006.

[28] A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, Multi-Class Clustering of Cancer Subtypes through SVM Based Ensemble of Pareto-Optimal Solutions for Gene Marker Identification, PLoS One, vol. 5, no. 11, pp. e13803, 2010.

[29] S. Ray, A. Mukhopadhyay and U. Maulik, Predicting Annotated HIV-1Human PPIs using a Biclustering Approach to Association Rule Mining, In Proc. EAIT-2012, pp. 28-31, Kolkata, India, November 2012.

[30] R. Raji, F. Guzzo, L. Carrara, J. Varughese, E. Cocco, S. Bellone, M. Betti, P. Todeschini, S. Gasparrini, E. Ratner, D.A. Silasi, M. Azodi, P. Schwartz, T.J. Rutherford, N. Buza, S. Pecorelli and A.D. Santin, Uterine and ovarian carcinosarcomas overexpressing Trop-2 are sensitive to hRS7, a humanized anti-Trop-2 antibody, Journal of Experimental & Clinical Cancer Research, vol. 30, no. 106, 2011.

[31] M. Hahsler, C. Buchta, B. Gruen and K. Hornik, Package ‘arules’, 2013, http://R-Forge.R-project.org/projects/arules/.

[32] T. Jayalakshmi and A. Santhakumaran, Statistical normalization and back propagation for classification, International Journal of Computer Theory and Engineering, vol. 3, no. 1, pp. 1793-8201, 2011.

[33] M. Bibikova and J. B. Fan, GoldenGate Assay for DNA Methylation Profiling, Methods in Molecular Biology, vol. 507, pp. 149-163, 2009.