International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Boolean Rule Based Classification for Microarray Gene Expression Data

R. Vengatesh Kumar, R. Lawrance

Abstract: ***Microarray technology provides a way to identify the expression level of ten thousands of genes simultaneously. This is useful for prediction and decision for the cancer treatments. To analyze and classify the gene expression data is more complex task. The rule based classifications are used to simplify the task of classifying genes. In this paper, a novel Boolean Rule based Classification (BRC) algorithm has been proposed. The efficient and relevant Boolean rules are assisting in classifying the test data correctly by Boolean Rule based Classifier model. This model is useful for drug designers. The experimental results show that in many cases the Boolean rule based classification yields more accurate results than other classical approaches.***

Keywords: Gene Expression, Gene selection, k-Means, Discretization, Boolean Rule Based Classification.

I. INTRODUCTION

DNA microarrays can measure the expression level of thousands of genes within a particular messengerRNA sample by simultaneously [4]. Microarray technology helps to identify the gene functions and expression levels. The causes of cancer gene expression data have been focused by the research community. Many research works investigate several aspects of cancer gene expression that includes exploration of the interactions between the abnormal and normal genes.

Classification is a supervised technique, which learns to classify new instances based on the knowledge learnt from a previously classified training set of instances. It takes a set of data already divided into predefined groups and searches for patterns in the data. A large number of classification algorithms have been developed and implemented to extract information and discover the knowledge patterns for decision making. Typical Classification Algorithms are Random Forest, Neural Networks, k-Nearest Neighbor and Bayesian Networks [3]. Recent experimental research indicates that rule based classification build good classifier models in terms of predictive accuracy when compared to other classification methods such as Bayesian Network, Neural Network and Random Forest [5].

Manuscript published on 30 September 2019
* Correspondence Author
R.VengateshKumar*, Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India.
R.Lawrance, Director, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

Rule based classification is an emerging classification research topic which employs association rules to solve classification problems in data mining. Rule based classification which integrates association rules and classification approach. The rule based classification algorithm uses the learned rules to predict the right class labels for each test data. It plays vital role for the proposed classifier model. The major works related to statistical testing, discretization and classification methods in biological data are described below.Jeanmougin *et al*. have stated that the statistical approaches are very important in biological research to identify the differentially expressed genes. They have conducted a survey on gene expression data using various statistical approaches, such as, Welch's t-test, Analysis of Variance(ANOVA), Significance Analysis of Microarrays (SAM),etc. The main statistical approach is to select genes differentially expressed between two groups are to apply t-test [4].Cristian *et al.* have discussed that the discretization is a technique frequently applied as a preprocessing step in the analysis of biological data. Discretization is to allow the application of algorithms for the biological knowledge that requires discrete data as an input, by mapping the real data into a typically small number of finite values. The biological problems are addressed by discretizing the gene expression data [1].

Wur *et al.* have proposed effective Boolean algorithm for mining association rules in market basket data analysis. This approach reduces the number of scans to generate the frequent item sets over the Apriori algorithm [11]. Mahajan *et al.* have discussed and evaluated the Rule Based Classification Algorithms. The rule based classification algorithms are experimentally compared with number of classified instances, accuracy and error rate using WEKA tool [7].

Thangaraj *et al.* have discussed different types of classification algorithms like tree-based, rule-based methods. They have compared and analyzed the performance of different rule based classifiers across multiple database relations [10]. Qin *et al*, have proposed a new rule based classification and prediction algorithm called *uRule* for classifying uncertain data. The algorithm introduces new measures for generating, pruning and optimizing rules [8].

Koklu *et al.* have proposed a novel application of rule based classification technique. They have stated that, traditional methods can't cope with turning the data to the knowledge. Classification and associated rule extraction of data mining techniques are preferred widely [5]. Suzan *et al.* have stated that choosing the right rules in the classifier to assign the predicted class is a challenging task.

Predicting the right class for each test data is vital because the overall predictive performance of the classifier depends on this decision [9]. From the literature survey, the statistical method has to be used to select the informative genes. The microarray data have to be discretized for rule generation. The existing traditional classification algorithms are tedious task to interpret the interesting biological knowledge from gene expression data. Rule based classifications are very easy to interpret the interesting knowledge. The rule based classification algorithm use association rules to predict the class labels of the test data. Hence, in this paper, a new algorithm called Boolean Rule based Classification (BRC) for Microarray Gene Expression Data is designed to classify the gene expression intervals data using association rules.

The remaining part of the paper is organized as follows. The proposed methodology of Boolean rule based classification for gene expression data is illustrated in Section II. The experimental outcomes and comparative analysis are shown in Section III. Conclusion and future work of this research is discussed in Section IV.

II. METHODOLOGY

Data mining techniques are frequently used to form a classifier that determines belonging class of a new data among the predetermined classes. The following phases are involved in the development of the proposed BRC Model for classifying breast cancer2 type gene expression data. The block diagram of the BRC Model is depicted in Fig.1 and its algorithm is illustrated in Fig. 2.

2.1 Data Collection

Breast cancer is the type of cancer with the highest incidence rate among women globally and it ranks the first among reasons for death due to cancer in women. The microarray breast cancer2type gene expression data set were taken from National Centre for Biotechnology Information (NCBI) [12].The dataset of 60 patients with ER-positive primary breast cancer data were generated from whole tissue sections of breast cancers. The sample breast cancer2 type gene expression data with two class labels of normal and abnormal which are shown in table 1. The rows represent the genes and columns represent the samples.

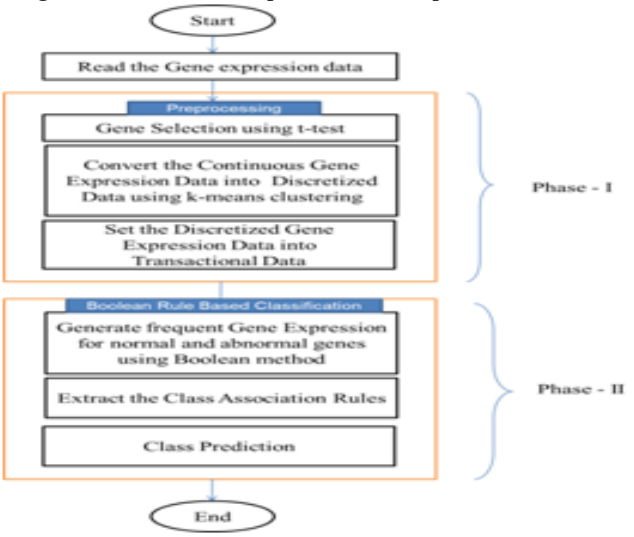

Fig.1 Block diagram of Boolean Rule Based Classifier Model

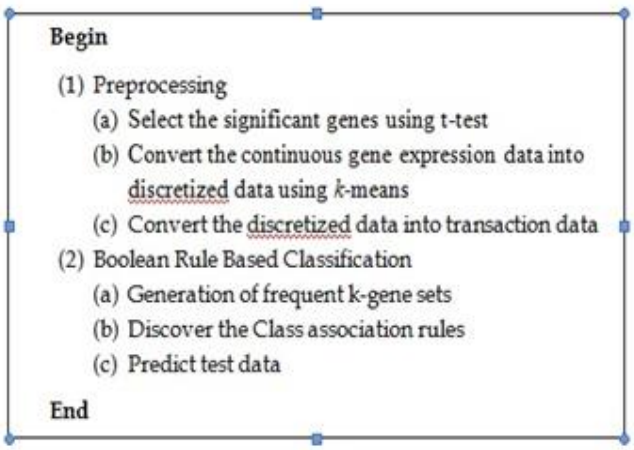

Fig.2 Boolean Rule based Classification (BRC) Algorithm

Table-1: Sample Gene Expression Data

Sample	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Class label	abnormal	abnormal	abnormal	abnormal	abnormal	normal	normal	normal	normal	normal
LYPD6	0.45	-1.49	-0.46	0.65	0.08	1.35	1.73	0.49	3.32	2.43
PTGER3	1.66	2.26	1.22	3.98	3.59	2.68	5.27	3.87	2.56	3.06
EST_1	-0.62	-0.68	-0.8	-0.68	-0.83	-0.09	-0.34	-0.61	-0.94	-0.52
EST_2	-0.49	-0.38	-0.56	-0.48	-0.41	0.04	-0.26	-0.07	-0.24	0.32
CHDH	0.86	1.17	0.34	1	-1.82	0.55	2.51	2.04	0.9	3.38
EST_3	0.64	0.48	-0.34	-0.04	-1.24	1.12	1.89	2.11	0.95	2.34
IL17BR	-0.7	-1.1	-2.16	-0.59	-3.56	0.76	1.47	1.93	-1.21	1.12
SCYA4	7.13	7.11	7.05	6.51	8.58	6.51	7.19	6.22	7.49	6.66
IL1R2	1.37	1.65	0.53	1.44	1.1	0.53	0.34	0.45	0.02	-0.14
ABCC11	5.96	6.55	4.35	6.82	6.02	3.67	0.36	0.28	1.19	1.24
HOXB13	-3.1	3.2	2.05	-3.33	1.12	0.01	-2.39	-3.34	-2.83	-0.79
APS	-1.45	0.68	-0.17	0.22	-0.45	0.17	-0.78	-0.63	-0.85	-0.75
ESTs_4	-2.57	2.3	1.11	-2.54	0.59	-0.45	-1.87	-2.62	-2.22	-2
DOK2	2.04	1.69	1.77	3.09	2.24	2.28	1.04	2.15	2.53	1.04
EST_5	1.83	1.55	1.94	1.25	2.1	1.65	1.3	1.2	1.67	0.87
GUCY2D	1.99	4.25	5.49	1.56	2.59	2.58	3.89	3.51	0.73	1.4

2.2 Preprocessing

The raw microarray gene expression data is often inconsistent. Preprocessing is one of the data mining techniques which involve transforming raw microarray data into understandable data. Data preprocessing prepares raw microarray data for further processing by using data selection and transformation. The data selection is a technique that is applied to select the significant genes expression data. The data discretization is a technique that is applied to transform the data from continuous into discrete for mining process.

a. Gene Selection

In this paper, the significantly expressed genes are selected using t-test. The BRC algorithm use t-test to provide estimates of significantly expressed genes by *p*-value. The *p*-value is used to determine the number of genes which are significantly different from normal gene expression data. The significant genes are chosen based on probability $p < 0.05$ and the insignificant genes are removed from gene expression [4].

In this experiment 600 significant genes are selected from 22575 genes. The table 2 represents the filtered microarray gene expression data.

Table 2: Filtered Gene Expression Data

Sample	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Class label	abnormal	abnormal	abnormal	abnormal	abnormal	normal	normal	normal	normal	normal
LYPD6	0.45	-1.49	-0.46	0.65	0.08	1.35	1.73	0.49	3.32	2.43
EST_2	-0.49	-0.38	-0.56	-0.48	-0.41	0.04	-0.26	-0.07	-0.24	0.32
EST_3	0.64	0.48	-0.34	-0.04	-1.24	1.12	1.89	2.11	0.95	2.34
IL17BR	-0.7	-1.1	-2.16	-0.59	-3.56	0.76	1.47	1.93	-1.21	1.12
IL1R2	1.37	1.65	0.53	1.44	1.1	0.53	0.34	0.45	0.02	-0.14
ABCC11	5.96	6.55	4.35	6.82	6.02	3.67	0.36	0.28	1.19	1.24

b. Gene Data Discretization

Data discretization is commonly used as pre-processing technique that transforms the number of continuous data by dividing its range into a finite set of distinct intervals [1,2]. Later, the data are analyzed in the knowledge representation and the discrete values are used to simplified data representation in the data exploration and mining process. In this paper, the filtered continuous microarray gene expression data are converted into discrete gene expression intervals data using k-means discretization algorithm. The table 3 depicts the discretized gene expression intervals data.

Table- 3: Discretized Gene Expression Data

Sample	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Class label	abnormal	abnormal	abnormal	abnormal	abnormal	normal	normal	normal	normal	normal
LYPD6	-0.29: 0.65	-1.49: -0.29	-1.49: -0.29	-0.29: 0.65	-0.29: 0.65	0.49: 2.03	0.49: 2.03	0.49: 2.03	2.03: 3.32	2.03: 3.32
EST_2	-0.56: -0.45	-0.45: -0.38	-0.56: -0.45	-0.56: -0.45	-0.45: -0.38	-0.05: 0.32	-0.26: -0.05	-0.26: -0.05	-0.26: -0.05	-0.00: 0.32
EST_3	-0.52: 0.64	-0.52: 0.64	-0.52: 0.64	-0.52: 0.64	-1.24: -0.52	0.95: 1.57	1.57: 2.34	1.57: 2.34	0.95: 1.57	1.57: 2.34
IL17BR	-1.83: -0.59	-1.83: -0.59	-3.56: -1.83	-1.83: -0.59	-3.56: -1.83	1.57: 2.34	1.57: 2.34	1.57: 2.34	-1.21: -0.55	1.57: 2.34
IL1R2	0.96: 1.65	0.96: 1.65	0.53 : 0.96	0.96: 1.65	0.96: 1.65	0.19: 0.53	0.19: 0.53	0.19: 0.53	-0.14: 0.19	-0.14: 0.19
ABCC11	5.34: 6.82	5.34: 6.82	4.35: 5.34	5.34: 6.82	5.34: 6.82	2.22: 3.67	0.28: 2.22	0.28: 2.22	0.28: 2.22	0.28: 2.22

c. Transaction Data

Association rule mining generates frequent item sets and discovers association rules using transaction dataset. Hence the discretized gene expression intervals data are converted into transaction data by considering samples as transactions and genes as transaction items, where the transaction data set are generated separately for normal and abnormal classes. The table 4 depicts the transaction data set for abnormal class. The table 5 depicts the transaction data set for normal class. These transaction data are used to generate **the class association rules.**

Table- 4: Transaction Data set for abnormal class

tID	Items
S1	LYPD6 [-0.29: 0.65] EST_2 [-0.56:-0.45] EST_3 [-0.53: 0.65] IL17BR [-1.83,-0.59] IL1R2 [0.96,1.65] ABCC11 [5.34,6.82]
S2	LYPD6 [-1.49:-0.29] EST_2 [-0.45:-0.38] EST_3 [-0.53: 0.65] IL17BR [-1.83:-0.59] IL1R2 [0.96:1.65] ABCC11 [5.34:6.82]
S3	LYPD6 [-1.49:-0.29] EST_2 [-0.56:-0.45] EST_3 [-0.53: 0.65] IL17BR [-3.56:-1.83] IL1R2 [0.53:0.96] ABCC11 [4.35: 5.34]
S4	LYPD6 [-0.29: 0.65] EST_2 [-0.56:-0.45] EST_3 [-0.53: 0.65] IL17BR [-1.83:-0.59] IL1R2 [0.96:1.65] ABCC11 [5.34:6.82]
S5	LYPD6 [-0.29: 0.65] EST_2 [-0.45:-0.38] EST_3 [-1.24 : -0.53] IL17BR [-3.56:-1.83] IL1R2 [0.96:1.65] ABCC11 [5.34:6.82]

Table- 5: Transaction Data set for normal class

tID	Items
S6	LYPD6 [0.49:2.03] EST_2 [0.00: 0.32]EST_3 [0.95:1.57] IL17BR [0.05: 1.93]IL1R2 [0.19: 0.53] ABCC11 [2.22:3.67]
S7	LYPD6 [0.49:2.03] EST_2 [-0.26:0.00]EST_3 [1.57:2.34] IL17BR [0.05: 1.93]IL1R2 [0.19: 0.53] ABCC11 [0.28:2.22]
S8	LYPD6 [0.49:2.03] EST_2 [-0.26:0.00]EST_3 [1.57:2.34] IL17BR [0.05: 1.93]IL1R2 [0.19: 0.53] ABCC11 [0.28:2.22]
S9	LYPD6 [2.03:3.32] EST_2 [-0.26:0.00]EST_3 [0.95:1.57] IL17BR [-1.21: 0.05]IL1R2 [-0.14: 0.19] ABCC11 [0.28:2.22]
S10	LYPD6 [2.03:3.32] EST_2 [0.00: 0.32]EST_3 [1.57:2.34] L17BR [0.05: 1.93]IL1R2 [-0.14: 0.19] ABCC11 [0.28:2.22]

2.3 Boolean Rule Based Classification

The BRC model is built to classify the gene expression using Boolean Class Association Rules. The Boolean method generates the frequent microarray genes expression without generating the candidate gene sets and extracting the class association rules in two steps.

Step-1: Frequent gene sets are identified using Boolean OR and AND operations.

Step-2: Class association rules are generated using Boolean AND and XOR operations

a. Frequent gene sets

Given a set of genes D= $\{g_1, g_2, g_3 \ldots g_n\}$ and a set of samples tID = $\{s_1, s_2, s_3 \ldots s_m\}$, a subset of D, S⊆D is called a frequent gene sets. The support(S) ≥ minimum support, where minimum support is a user defined threshold value [3, 6, 11].

The transaction data are used to mine the frequent gene set for normal and abnormal classes. The table 6 represents the normal frequent gene set. The table 7 represents the abnormal frequent gene set.

Table 6: Frequent gene set for normal class

S.No	Frequent Geneset
1	LYPD6[0.49 : 2.03], IL17BR[0.05 : 1.93],IL1R2[0.19 : 0.53]
2	EST_3[1.57 : 2.34], IL17BR[0.05 : 1.93],ABCC11[0.28 : 2.22]

Table 7: Frequent gene set for abnormal class

S.No	Frequent Geneset
1	EST_2[-0.56 : -0.45], IL17BR[-1.83 : -0.59], EST_3[-0.527 : 0.640]
2	LYPD6[-0.29 : 0.65], IL1R2[0.96 : 1.65], ABCC11[5.34 : 6.82]
3	EST_3[-0.52 : 0.64],IL17BR[-1.83 : -0.59], IL1R2[0.96 : 1.65]
4	EST_3[-0.52 : 0.64],IL17BR[-1.83 : -0.59], ABCC11[5.34 : 6.82]

b. Class Association Rules

Given a set of Class Association rule, let $R = \{r_1, r_2, r_3 \dots r_n\}$ is a rule set where n represents the number of rules, $C = \{c_1, c_2, c_3 \dots c_m\}$ is a class label where m represents the number of class labels and $A = \{a_1, a_2, a_3 \dots a_m\}$ is a set of all items that appear in D. A class association rule is defined in the form of $A \longrightarrow C_m$. The left-hand side of the rule is said to be antecedent and right-side of the rule is said to be class label. The table 8 represents the class association rules which are known as Boolean rule based classification model based on the minimum support 50% and minimum confidence 75%.

Table 8: Class Association Rules

Gene Sample	Rule No	Class Association Rules	
G1	R1	LYPD6[0.49 : 2.03] , IL1R2[0.19 : 0.53]	→Normal
	R2	EST_3[1.57 : 2.34] , IL17BR[0.06 : 1.93	→Normal
	R3	EST_3[1.57 : 2.34] , IL17BR[0.05 : 1.93] , ABCC11[0.28 : 2.22]	→Normal
G2	R4	EST_2[-0.56 : -0.45], EST_3[-0.53 : 0.64]	→Abnormal
	R5	LYPD6[-0.29 : 0.65] , IL1R2[0.96 : 1.65]	→Abnormal
	R6	IL17BR[-1.83 : -0.59], EST_3[-0.53 : 0.64]	→Abnormal
	R7	EST_3[-0.53 : 0.64], IL17BR[-1.83 : -0.59] , ABCC11[5.34 : 6.82]	→Abnormal

c. Class Prediction

The test data has been predicted using the Boolean rule based classifier. A rule r covers a record t, if the attributes of record satisfy the condition of the rule then class label C_m assign to rule r, if none of the gene attribute is not presented in the rule then assign the class label as default class in the training data set. $A \longrightarrow C_m$, where A is the conjunctions of attributes, C_m is the class label.

Table 9 represents the test data. The class label prediction algorithm is illustrated in Fig. 3, which is used to predict the class label and evaluate the classifier model.

Table 9: Sample Test Data

Sample	LYPD6	EST_2	EST_3	IL17BR	IL1R2	ABCC11	Class in Training Data
T1	0.45	-0.49	0.64	-0.7	1.37	5.96	abnormal
T2	1.35	0.04	1.12	0.76	0.53	3.67	normal
T3	-1.49	-0.38	0.48	-1.1	1.65	6.55	abnormal
T4	3.32	-0.24	0.95	-1.21	0.02	1.19	abnormal
T5	1.73	-0.26	1.39	1.47	0.34	0.36	normal
T6	-0.46	-0.56	-0.34	-2.16	0.53	4.35	abnormal
T7	0.65	-0.48	-0.04	-0.59	1.44	6.82	abnormal
T8	0.49	-0.07	2.11	1.93	0.45	0.28	normal
T9	0.08	-0.41	-1.24	-3.56	1.1	6.02	abnormal
T10	2.43	0.32	2.34	1.12	-0.14	1.24	normal

Input: test data, D
Output: Class label Prediction
Procedure **Class-Prediction** (Class Rules R, Test A)
Begin
Initialize flag=0, count=0, count1=0
(a) Read Test data $A = (a_1, a_2, a_3 \dots a_n)$
(b) Read Class Rules $R = (r_1, r_2, r_3 \dots r_n)$ of group, G1- normal and G2 – abnormal
(c) for($a = 1$:$a \le n$; $a = a + 1$)
{
for($G_1 = 1$:$G_1 \neq \emptyset$; $G_1 = G_1 + 1$)
if attribute a_i present and covers rule in G_1 *then*
count =count +1 ;flag=1;
for($G_2 = 1$:$G_2 \neq \emptyset$; $G_2 = G_2 + 1$)
if attribute a_i present and covers rule in G_2 *then*
$count_1$=$count_1$+1 ;flag=1;
}
(d) *if* (flag==1)
Test data attributes cover rules
if (count <= $count_1$) *then*
assign the class label as abnormal
else
assign the class label as normal
end if
else
Test data attributes does not cover any rule,
assign the default class label as normal
end if
end

Fig.3 Class label Prediction Algorithm

III. EXPERIMENTAL RESULTS

The BRC model checks each attribute of the test data with class association rules. The test data of any one attribute is present and covers the rules the corresponding class label count is incremented. After the completion of each attribute of the test data, the number of occurrences of the normal and abnormal class label is counted. Finally the class labels occurrences are compared and assign class label to the test data which has highest count. The abnormal class label is assigned to the test data when the count of both normal and abnormal class label is equal. The table 10 represents the number of class occurrences of class results. The test data of all the attributes doesn't cover any rules assign the default class label as normal.

Table 10: Classified Results

	Class Label Occurrences		
Sample	Attribute covers in normal class	Attribute covers in abnormal class	Predicted Class
T1	0	9	abnormal
T2	4	0	normal
T3	0	7	abnormal
T4	1	2	abnormal
T5	5	0	normal
T6	1	5	abnormal
T7	1	6	abnormal
T8	7	1	normal
T9	0	3	abnormal
T10	5	0	normal

The Fig. 4 shows that BRC Model gives the best performance comparing to other algorithms. The OneR, Random Forest, Naïve Bayesian and KNN classification algorithm has the average accuracy performance. BRC Model has the minimum error rate as compare to other studied algorithms and also it has overall performance.

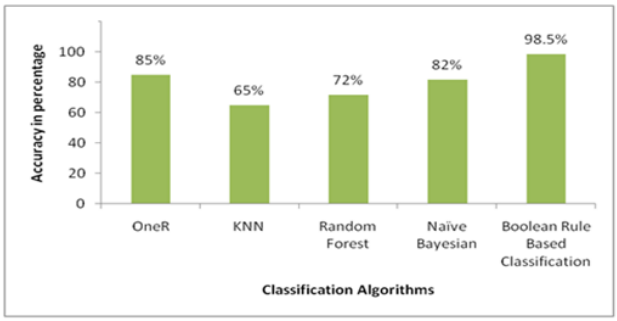

Fig. 4 Comparative Analysis of Various Classification Algorithms

IV. CONCLUSION

The Boolean rule based classification algorithm has experimented by using the microarray breast cancer2 type dataset. The experimental results show that the BRC model gives the high accuracy and low error rate to compare with other traditional algorithms and also it assures that the best class assignment for each test case in the gene expression data.

The proposed classifier model is used to classify the diseased gene expressions like LYPD6, IL1R2, ABCC11, EST_2, EST_3which provides the gene targeting treatments. In future, the BRC algorithm will be used to analyze the various biological data.

REFERENCES

1. Cristian A.G, Rocio L.C, Jessica A.C,Micheletto S,Ponzoni I. Discretization of gene expression data revised, Briefings in Bioinformatics, 2016; 17(5), 758–770.
2. Garcia S, Luengo J, Sáez J.A, Herrera F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning, Knowledge and Data Engineering, IEEE Transactions, 2013; 25(4),734-750.
3. Han J, Kamber M. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Elsevier, 2002.
4. Jeanmougin M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies,PloS one 2010; 5(9).
5. Koklu M, Kahramanli H, Allahverdi, N. Applications of rule based classification techniques for thoracic surgery. In the Make Learn and TIIM Joint International Conference on Managing Intellectual Capital and Innovation for Sustainable and Inclusive Society: Managing Intellectual Capital and Innovation, ToKnowPress, Bari, Italy, 1991.
6. Martinez R, Nicolas P, Claude P. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations, Bioinformatics, 2008; 24(22),2643-2644.
7. Mahajan A, Ganpati A. Performance evaluation of rule based classification algorithms, International Journal of Advanced Research in Computer Engineering & Technology, 2014;3546-3550.
8. Qin B, Xia Y, Prabhakar S, Tu Y. A rule-based classification algorithm for uncertain data in Data Engineering, ICDE'09, 2009; 1633-1640.
9. Suzan A, Joan L, Fadi T. Class Strength Prediction Method for Associative Classification, The Fourth International Conference on Advances in Information Mining and Management,2014.
10. Thangaraj M, Vijayalakshmi C. R. Performance Study on Rule-based Classification Techniques across Multiple Database Relations, International Journal of Applied Information Systems, 2013; 2249-0868
11. Wur S.Y, Leu Y. An Effective Boolean Algorithm for Mining Association Rules in Large Databases, Database Systems for Advanced Applications, IEEE Transactions, 1999;179-186.
12. www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379

AUTHORS PROFILE

R.VengateshKumar, Research Scholar, Research & Development centre,
Bharathiar University, Coimbatore, Tamil Nadu, India.

R.Lawrance, Director, Department of Computer Applications,
Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India