CLASSIFICATION OF SAMPLE CONTAINING GENE EXPRESSIONS INTO DISEASED AND NORMAL CLASS

[1]Gauri Bhanegaonkar, [2]Rakhi Wajgi, [3]Dipak Wajgi
[1] PG scholar, [2,3] Professor Computer Science and Engineering
[1] Department of Computer Science and Engineering,
[1]Yeshwantrao College of Engineering, Nagpur, India

Abstract : Data mining is a new dominant technology with a great potential to discover useful information. Associative classification is a recent appeal in data mining that promote association rule discovery and to build classification model. The main objective of this task is to classify sample containing gene expressions into cancer and normal class by using association classification. The designed algorithm contains four stages called as Statistical Gene Filtering, Binary Discretization, Class Association Rules and Class Prediction. The gene filtering is to search the differentially expressed genes and favored the significant gene in the specific gene expression. The binary discretization is to splits the continuous values into two intervals such as 0 or 1. The class association rules are generated by using closed frequent itemset and build the classifier model. The last stage is to predict the class from trained classifier model by using scoring function. The database is given from NCBI online biological database.

***IndexTerms* – Data mining, microarray data, Associative classification.**

I. INTRODUCTION

The vast spread of electronic data collection in medical surroundings leads to an aggressive growth of clinical data derived from heterogeneous patient samples. Collecting, maintaining, coordinating and analyzing these data are crucial actions in order to shed light on disease such as cancer and on related therapy. Data mining is the mechanism of discovering anomalies, patterns and correlations between genes in large data sets to predict the class.

The field of data mining has seen huge success in terms of wide ranging application. In this paper, the prospective methodology targets on associative classification which is a advanced approach in the area of data mining for classifying the biological data in the field of bioinformatics. Data mining is the discipline of discovering interesting patterns and relationships in ample volume of data. This process is also called as Knowledge Discovery in Databases.

A microarray database is storehouse consists of gene expression data. Microarray dataset is used to measure the expression levels of large amount of genes concurrently. The microarray database is represented in the form of M x N matrix of gene expression values, where row shows genes g1, g2, …., gn and column shows sample s1, s2,…., sn. The figure 1 represents the matrix of gene expressions.

P	G1	G2	G3	...	Gm	Class
S1	G(1,1)	G(1,2)	G(1,3)	…	G(1,n)	Cancer
S2	G(2,1)	G(2,2)	G(2,3)	…	G(2,n)	Normal
S3	G(3,1)	G(3,2)	G(3,3)	…	G(3,n)	Normal
...	…	…	…	…		
Sn	G(n,1)	G(n,2)	G(n,3)	…	G(n,m)	Cancer

Fig. 1. Microarray Gene Expression Data

The microarray breast cancer dataset is downloaded from NCBI biological database with series GSE1379 training dataset and GSE1380 for testing dataset. Training dataset comprise of 60 samples, out of which 32 samples exists as cancer and 28 samples exists as normal.

II. METHODOLOGY

Microarray technology has become an effective approach in the analysis of gene expression for the recognition of disease genes and therapeutic targets for human cancers. But microarray produces vast amount data to the researcher. At the same time, it is difficult to acquire accurate and understandable knowledge from the data. So the computer programs can automatically finds the affected genes and accurately classify the gene expression.

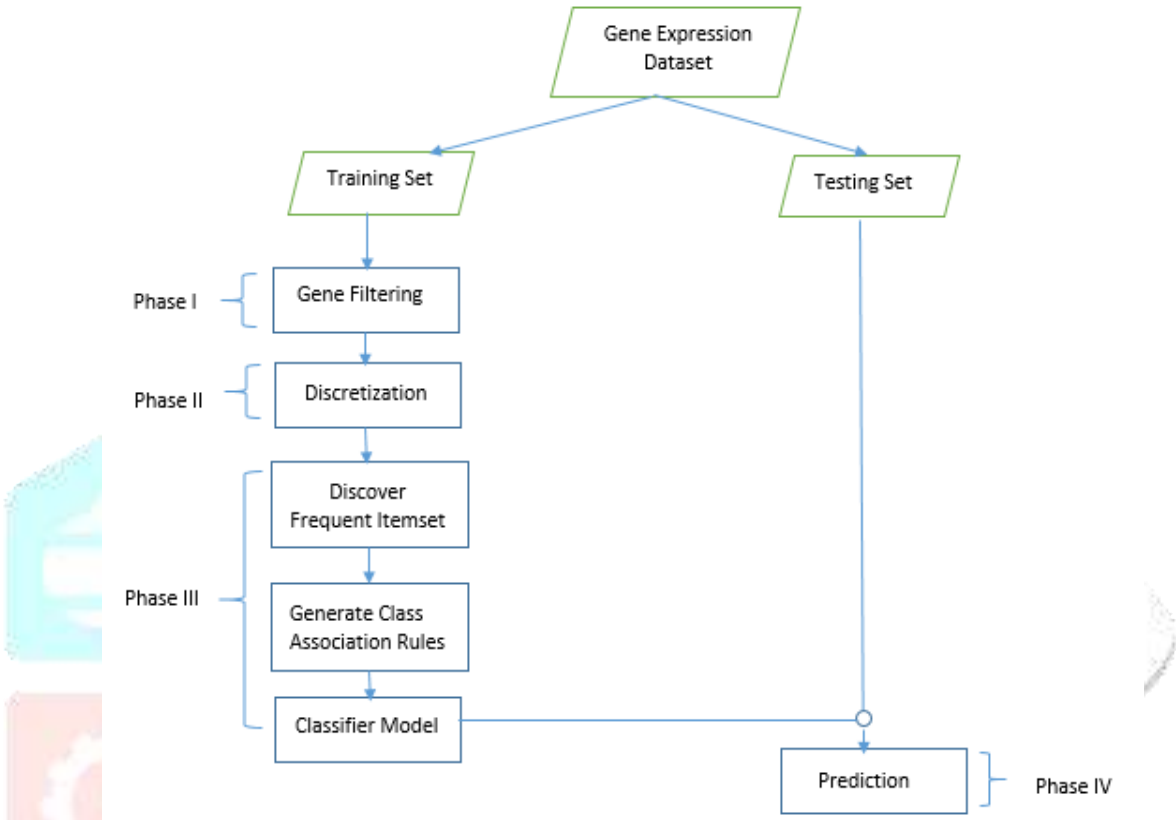

Fig 2: Methodology

Phase I: Gene Filter

The human body cell mostly expresses enormous volume of genes. The modification in expression patterns of few genes is acceptable to change the subsequent phenotype of cell, for that sense hypothesis testing is used to explain the genetic divergence between cells. Microarray technology allows the analysis of thousands of genes simultaneously. Gene filtering apliances a method called statistical independent t-test. The independent t-statistic applies by using following formula

$$T = \frac{(\bar{X} - \bar{Y})/\hat{\sigma}_{pool}}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{----------} \quad (1)$$

$$\sigma_{pool} = \frac{\sum_1^n X_i^2 - nX^{-2} + \sum_1^m Y_i^2 - mY^{-2}}{n+m-2} \quad \text{----------} \quad (2)$$

Where $\bar{X} - \bar{Y}$ represents is sample means and σ_{pool} is an unbiased estimator for standard deviation. The dissimilarity between normal and cancer cell is calculated using p value. In this paper, first the means and standard deviation for both groups are calculated. Then the t-statistic value and p value will be calculated by using t-statistic formula. Finally, if $p < 0.5$ then that gene is significant with degree of freedom n-2 and $p < 0.5$. The last phase of gene filtering removes the uninformative genes. 2339 genes are filtered out for further processing. Following figure shows the names of few filtered genes out of 22575.

Phase II: Discretization

Discretization process is one of the essential data preprocessing task in data mining. Discretization process consists of four steps such as, sorting the continuous values of the feature, calculating split points, splitting intervals of continuous value and finally stopping at some point. Discretization can be categorized into supervised or unsupervised depending upon the dataset. In this paper, supervised discretization preformed with taking class information. ID3 algorithm is used for discretization process of continuous features in the gene expression. The two parts of ID3 algorithm are entropy and information gain. The entropy is used to calculate the homogeneity of sample and the information gain measures how much information features gives us about class. The formula used for entropy is

$$H(X) = - \sum_{i=0}^{n} p_i * \log_2(p_i) \quad \text{----------} \quad 3$$

Where H(X) be the entropy. p_i specifies the probabilities of information X. ID3 applies a greedy search to catch split points in the period of existing range of continuous value using formula 4. In formula 4, m be the number of classes in X, , $p_{j,left}$ and $p_{j,right}$ are probabilities.

$$Info(S,T) = -p_{left} \sum_{j=1}^{m} p_{j,left} \log_2 p_{j,left}$$
$$-p_{left} \sum_{j=1}^{m} p_{j,right} \log_2 p_{j,right} \quad \text{----------} \quad 4$$

In this paper, binary discretization used to discretize continuous values to 0 or 1 form. Binary discretization transforms continuous values to discrete values such as 0 or 1. In proposed methodology, for each gene attribute, entropy value is calculated and split values are also calculated for each gene attribute. Consequently, the gene expression values are replaced by 0 or 1. After discretization, transaction table is generated for both cancer and normal. The transaction table consists of combination of genes with information gain.

Phase III: Class Association Rules

Association rule mining is a well searched method to find interesting relations between genes in large databases. Apriori algorithm is one of the most popular algorithm which is used to extract frequent itemsets from ample amount of database and getting the association rule for discovering the knowledge. There are two phases of association rule mining such as finding frequent closed itemsets and and generate association rules.

Frequent Closed Itemsets

The two properties of closed frequent itemsets are, 1. All sub-closed itemsets of frequent closed itemset are frequent, 2. All sub-closed itemsets of an infrequent closed itemset are also being infrequent. The closed itemset C is said to be frequent if the itemset satisfy user defined support.

Class Association Rule

In our proposed work, the process of class association rule consists of two steps (1) the outcome of set of rules that satisfy user defined support and confidence. (2) construction of classifier model from these association rules. In proposed methodology, Apriori algorithm is used for mining frequent itemsets and significant association rules. This algorithm operates on transactional database which consists of combination of genes. The overall algorithm can be divided into two phases. (1) Assign minimum support to search the frequent sets wirh k items in a dataset. (2) With the help of frequent k-itemsets find the frequent sets with k+1 ites by using self-join rule. The algorithm first loads the set of generator itemsets. Each iteration consists of three steps. First, to find the candidate frequent closed itemsets and support using closure property. Next, the infrequent itemsets shorten from the frequent closed candidate itemsets. Lastly, develop the valid frequent closed itemsets. The itemsets generate the set of rules for cancer and normal.

Phase IV: Prediction

An algorithm produces interesting relationships between gene expressions and classes. The generated rule that satisfy user defined support and confidence have been used to build classifier model. The final step in associative classification is prediction. In proposed methodology, score based function is used for class assignment for test data. Let $\bar{S}$ is an unknown sample and let $R^c = \{r_1^c \ldots r_{r_h^c}^c\}$ is the association rules for class c. For $\bar{S}$, calculates how many rules are satisfied in each rule R^c and assigns $\bar{S}$ to that class c whose rules are maximally satisfied using scoring function.

$$E(r_n^c, \bar{S}) = \frac{|\frac{r_n^c \cap \bar{S}}{r_n^c} \log |r_n^c|}{|R^c|} \quad \text{----------1}$$

$$E = \sum_{n=1}^{h_c} E(r_n^c, \bar{S}) \quad \text{---------- 2}$$

In prediction phase, first reads the test set and association rules for the class then score based function calculates the number of rules satisfied. The score value is calculated for each rule. Lastly, the score value is given to the class who maximally satisfy the test data. If no rules are relevant to the test data, the default class is assigned to that test data.

Classification Accuracy

Classification accuracy is the count of correct predictions made divided by the total count of predictions made. A confusion matrix is a way to represent the prediction results of a classifier. In confusion matrix, each cell comprise of predictions made by the classifier. Following table represents confusion matrix. True positives are the cases in which 23 cancer samples are correct predicted, true negatives are the cases that they don't have the disease, false positives are the cases that predicted yes but they don't have the disease also called as type I error, false negatives are the cases that are predicted no but they actually have the disease also called as type II error. Sensitivity is also called as true positive rate and it is computed as

Sensitivity = TP / (TP + FN)

Specificity is also called as true negative rate and it is computed as

Specificity = TN / (TN + FP)

Type I Error

Confusion Matrix		Target		
		+	-	
Actual	+	TP (23)	FP (5)	
	-	FN (1)	TN (31)	
		Sensitivity	Specificity	
		23/(23+1) =0.95	31/(5+31) =0.86	Accuracy= (23+31)/(23+5+1+ 31)=0.9=90%

Type II Error

Fig 3: Confusion Matrix

Result and Discussion

A microarray used to expose the thousands of gene expressions at the same time. The microarray dataset of breast cancer has been collected from National Centre for Bioinformatics (NCBI) with series GSE1379. The original dataset comprise of 60 samples, 32 are cancer samples and 28 are normal samples. Each sample comprise of 22575 gene expressions. The following table represents the microarray gene expression data.

Table 1: Gene Expression Data

GeneName	**0**	**0**	**0**	**0**	**1**	**0**	**0**
(+)Pro25G	0.4273	-0.72281	-1.2059	-0.4206	-0.38161	-1.1094	-0.622
(+)Pro25G_onG3PDH570_1(	-1.0998	-0.0230	1.0162	-1.2836	-0.65343	-1.3678	-0.0495

60)							
(-)GD11	-0.042	0.2731	-0.2258	0.9602	0.076856	0.9031	-0.4559
(-)3xSLv1	0.03873	-0.0699	-0.060	-0.1790	0.61131	0.4251	0.0185
(+)Pro25G	1.315	0.2908	0.3058	0.4868	-0.327	0.2402	0.3386
(+)Pro25G_onG3PDH570_1(60)	0.33014	0.18493	0.19833	-0.3130	0.2081	0.117	0.1019
(-)GD11	0.38357	0.431105	0.31622	1.16963	0.45311	0.3865	0.2591
…	…	…	…	…	…	…	…

The independent t-test calculates t-value and p-value. The p-value is used to find if a number is significantly dissimilar from normal genes. These steps automatically runs while implementing t-statistic method. First determine t-statistic then from t value measure p value with n-2 degree of freedom. Finally, selecting significant genes with $p<0.05$. There are 2339 genes are filtered out from 22575 genes.

Similarly, the entropy and split values are calculated for each gene attribute in gene expression data. Consequently, gene expression values replaced with 0 or 1. The discretized gene expression values are presented in table 4.2

Table 2: Discretized Gene expression data

Gene_id	**Gene name**	**Samples**			
		GSM22449	GSM22450	GSM22451	GSM22452
26	(+)Pro25G_onG3PDH570_1(60)	1	0	1	1
43	(-)GD11	1	1	1	1
44	(-)3xSLv1	1	1	1	1
69	(+)Pro25G	1	0	1	1
73	(+)Pro25G	1	1	1	1
97	(+)Pro25G	1	1	1	1
…	…	…	…	…	…
22559	(-)GD11	1	1	1	1

Table 3: Cancer Transaction Table

Sr. No	**Gain**	**Combinations of gene IDs**	**Lower Bound**	**Upper Bound**	**True Count**	**Index**
1	-0.0992784	[(+)Pro25G, (+)Pro25G_onG3PDH570_1(60)]	[-2.6829,-0.05449]	[0.947321,0.77604]	28	[1,2]
2	-0.0992784	[(+)Pro25G, (-)GD11]	[-2.682965,-0.476589]	[0.947321,0.4549]	28	[1,3]

3	-0.099278	[(+)Pro25G, (+)Pro25G]	[-2.68296,0.087531]	[0.94732,0.7514	28	[1,4]
4	-0.099278	[(+)Pro25G,(+)Pro25G_onG3PDH570_1(60)]	[-2.6829653,-3.9652293]	[0.94732,-1.55360]	28	[1,5]
5	-0.099278	[(+)Pro25G, (-)3xSLv1]	[-2.6829653,-2.6311633]	[0.947321,1.303525]	28	[1,6]
6	-]0.099278	[BC016056,(-)3xSLv1]	[-2.6829653,-0.1083013]	[0.947321,0.4165]	28	[1,7]
7	-0.099278	[AK054839, (-)GD11]	[-2.6829653,-0.1651753]	[0.94732,1.10648]	28	[1,8]
8	-0.099278	[AK054839, (-)3xSLv1]	[-2.6829653,-0.1425033]	[0.94732,0.28807]	28	[1,9]
…	…	…	…	…	28	…
190	-0.14598	[(-)GD11, (-)3xSLv1]	[-0.1713953,-0.1989963]	[0.16247,0.36939]	28	[19,20]

Table 4: Normal Transaction Table

Sr. No	Gain	Combinations of gene IDs	Lower Bound	Upper Bound	True Count	Index
1	-0.09927	[(+)Pro25G, (+)Pro25G_onG3PDH570_1(60)]	[-1.925313,-0.115288]	[0.7246867,0.3747117]	28	[1,2]
2	-0.099278	[(+)Pro25G, (-)GD11]	[-1.925313,-0.325889]	[0.724686,0.59411]	28	[1,3]
3	-0.099278	[(+)Pro25G, (+)Pro25G]	[-1.925313,0.0613993]	[0.724686,0.511399]	28	[1,4]
4	-0.099278	[(+)Pro25G,(+)Pro25G_onG3PDH570_1(60)]	[-1.925313,-4.022678]	[0.724686,-2.952678]	28	[1,5]
5	-0.099278	[(+)Pro25G, (-)3xSLv1]	[-1.925313,-2.736358]	[0.7246867,2.093641]	28	[1,6]
6	-0.099278	[BC016056,(-)3xSLv1]	[-1.925313,-0.134175]	[0.7246867,0.3358247]	28	[1,7]
7	-0.099278	[AK054839, (-)GD11]	[-1.925313,-0.068235]	[0.7246867,1.1617647]	28	[1,8]
8	-0.099278	[AK054839, (-)3xSLv1]	[-1.925313,-0.137441]	[0.7246867,0.262558]	28	[1,9]
…	…	…	…	…	28	…
190	-0.145981	[(-)GD11, (-)3xSLv1]	[-0.247710,-0.063687]	[0.0822897,0.2763127]	28	[19,20]

The above transaction table is used for rule generation. The class association rules are developed based on closed frequent itemset with support and confidence. The antecedent of each rule accomplish the membership of class c in consequent. Total 190 rules are generated for cancer and 190 for normal from transaction table. Support count indicates how frequently that rule appears. Confidence indicates how often the rule has been found true. Lift is calculated by using observed support to the expected.

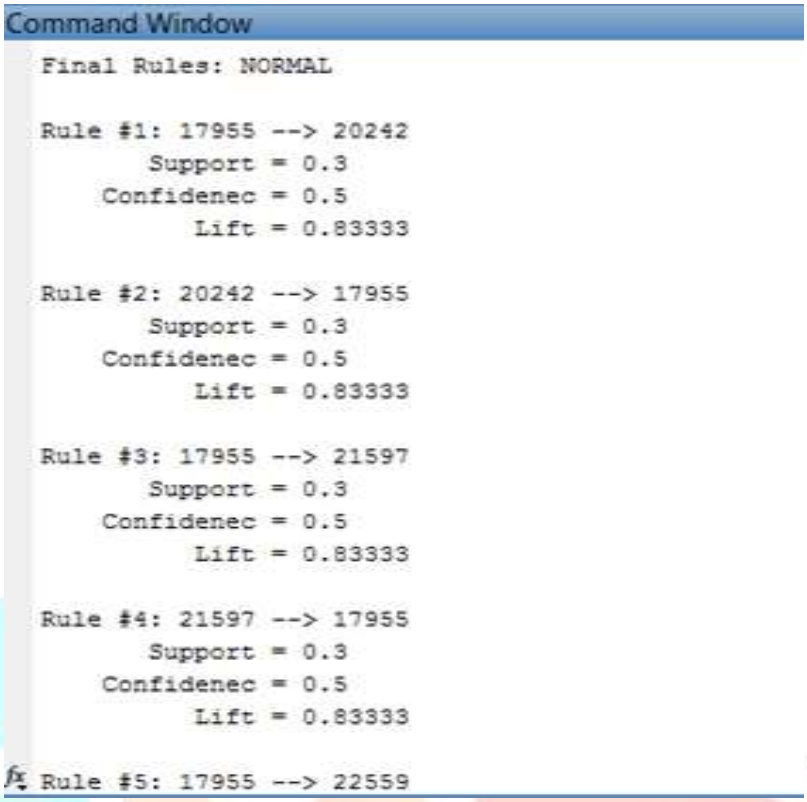

Fig 4: Rules for normal

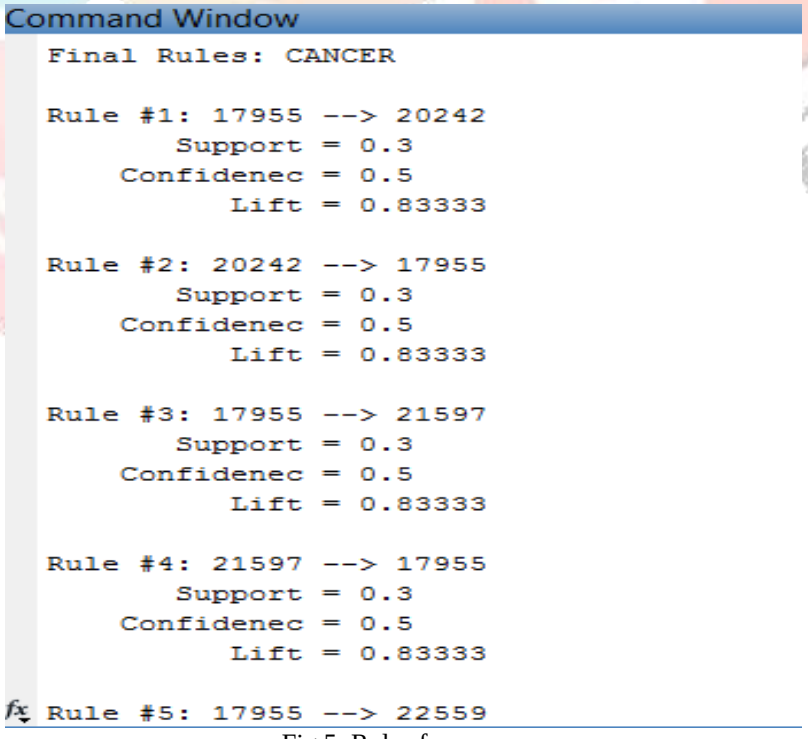

Fig 5: Rules for cancer

References:

[1] Rama Sreepada, Swati Vipsita, Puspanjali Mohapatra.,''An efficient approach for classification of gene expression microarray.'' In: Fourth International Conference of Emerging Applications of Information Technology, 2014.

[2] Girija Chetty, Madhu Chetty.: Multiclass Microarray Gene Expression Classification Based on Fusion of Correlation Features. In: IEEE Trans..

[3] Gerald Schaefer, Yasuyuki Yokota.: Fuzzy Classification of Gene Expression Data. In: IEEE 2007
[4] Ranjita Das, Sriparna Saha.: Gene expression classification using a fuzzy points symmetry based PSO clustering technique. In: Second International Conference on Soft Computing and Machine Intelligence,2015
[5] Ranjita Das, Sriparna Saha.: Microarray Gene Expression Data classification using Modified Differential Evolution Based Algorithm. In: IEEE INDICON 2015
[6] Benny Y. M. Fung, Vincent T. Y. Ng.: Classification of heterogeneous gene expression data. Volume 5, Issue 2
[7] Xi Hang Cao and Zoran Obradovic.: A Robust Data Scaling Algorithm for Gene Expression Classification. In: 2015 IEEE
[8] Kaimin Wu, Xiaofei Nan, Yumei Chai, Liming Wang,Kun Li.: DTSP-V:A Trend-based Top Scoring Pairs Method for Classification of Time Series Gene Expression Data. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
[9] Timothy M. Josserand.*:* classification of gene expression data using pca-based fault detection and identification. *In: 2008 IEEE*
[10] Su Liangliang, Wang Nian, Tang Jun, Chen & Le, Wang Ruiping.:The Classification of Gene Expression Profile Based on the Adjacency Matrix Spectral Decomposition. In: 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)
[11]Salvador Garcı´a, Julia´n Luengo, Jose´ Antonio Sa´ez, Victoria Lo´pez, and Francisco Herrera.: A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. In: IEEE transactions on knowledge and data engineering, vol. 25.
[12] Mansoor Raza, Iqbal Gondal, David Green, Ross L. Coppel.: Feature selection and classification of Gene expression profile in hereditary Breast cancer. In: Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04) IEEE
[13] Han,J. and Kamber,M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, *Elsevier*, 2002.
[14] Nagata, K., Washio, T., Kawahara, Y. and Unami, A.: Toxicity prediction from toxicogenomic data based on class association rule mining. *ELSEVIER journal*, Toxicology Reports, vol.41.