MINING OF IMPORTANT INFORMATIVE GENES AND CLASSIFIER CONSTRUCTION FOR CANCER DATASET

Soumen Kumar Pati[1] and Asit Kumar Das[2]

[1]Department of Computer Science/Information Technology, St. Thomas'College of Engineering and Technology, 4, D.H. Road, Kolkata-23
soumen_pati@rediffmail.com
[2]Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howrah-03
asitdas72@rediffmail.com

ABSTRACT

Microarray is a useful technique for measuring expression data of thousands or more of genes simultaneously. One of challenges in classification of cancer using high-dimensional gene expression data is to select a minimal number of relevant genes which can maximize classification accuracy. Because of the distinct characteristics inherent to specific cancerous gene expression profiles, developing flexible and robust gene identification methods is extremely fundamental. Many gene selection methods as well as their corresponding classifiers have been proposed. In the proposed method, a single gene with high class-discrimination capability is selected and classification rules are generated for cancer based on gene expression profiles. The method first computes importance factor of each gene of experimental cancer dataset by counting number of linguistic terms (defined in terms of different discreet quantity) with high class discrimination capability according to their depended degree of classes. Then initial important genes are selected according to high importance factor of each gene and form initial reduct. Then traditional k-means clustering algorithm is applied on each selected gene of initial reduct and compute miss-classification errors of individual genes. The final reduct is formed by selecting most important genes with respect to less miss-classification errors. Then a classifier is constructed based on decision rules induced by selected important genes (single) from training dataset to classify cancerous and non-cancerous samples of experimental test dataset. The proposed method test on four publicly available cancerous gene expression test dataset. In most of cases, accurate classifications outcomes are obtained by just using important (single) genes that are highly correlated with the pathogenesis cancer are identified. Also to prove the robustness of proposed method compares the outcomes (correctly classified instances) with some existing well known classifiers.

KEYWORDS

Microarray cancer data, K-means algorithm, Gene selection, Classification Rule, Cancer sample identification, Gene reducts.

1. INTRODUCTION

Now-a-days, an increasing number of applications in different fields especially on the field of natural and social sciences produce massive volumes of very high dimensional data under a variety of experimental constrains. In scientific databases like gene microarray dataset [1], it is common to encounter large sets of observations, represented by hundreds or more of dimensions. Microarray technology [2] allows to simultaneously analyzing thousands or more of genes and thus can give important insights about cell's function, since changes in the composition of an

DOI: 10.5121/ijsc.2012.3306

organism are generally associated with changes in gene expression patterns. The availability of massive amounts of experimental data based on genome-wide studies has given momentum in recent years to a large effort in developing mathematical, statistical, and computational techniques to surmise biological models from data. In many bioinformatics problems, number of genes is significantly larger than the number of samples (high gene-to-sample ratio data sets). This is typical of cancer classification tasks where a systematic investigation of the correlation of expression patterns of thousands of genes to specific phenotypic variations is expected to provide an improved catalog of cancer. In this context, the number of features corresponds to the number of expressed gene probes (up to several thousand) and the number of observations to the number of tumor samples (typically on the order of hundreds) is typically correlated.

In DNA microarray data [1] analysis generally biologists measure the expression levels of genes in the tissue samples from patients, and find explanations about how the genes of patients relate to the types of cancers they had. Many genes could strongly be correlated to a particular type of cancer, however, biologists prefer to focal point on a small subset of genes that dominates the outcomes before performing in-depth analysis and expensive experiments with a high dimensional dataset. Therefore, automated selection of the small subset of genes is highly advantageous. DNA microarray technology [2] has directed the focus of computational biology towards analytical data interpretation [3]. However, when examining microarray data, the size of the data sets and noise contained within the data sets compromises precise qualitative and quantitative analysis[4].

Generally, this field includes two key procedures: important gene identification and classifier construction. The gene selection [5,6] is particularly crucial in this topic as the number of genes irrelevant to classification may be huge, and hence, accurate prediction can be achieved only by performing gene selection reasonably, that is, identifying most informative genes from a large number of candidates. Once such genes are chosen, the creation of classifiers on the basis of the genes is another mission. Most of the papers [7-9] obtain accurate classification results based on more than two genes.

In the paper, a novel gene selection and subsequently a suitable classification rule generation technique has been proposed on microarray data for selecting a single important gene to predict cancerous gene with high classification accuracy. The method can be broken down into following four steps:

i. The gene expression dataset is standardized to Z-score using Transitional State Discrimination method [10] and then discretized to five discrete values.
ii. Since, all genes are not important to identification of particular cancer diseases, a relevance analysis of genes are performed to select only the important genes. As the samples of genes are collected from both normal and cancerous patients, the samples are divided into two disjoint classes. For each gene, frequencies of discrete sample values are computed in each class, based on which importance of the genes is measured.
iii. Since, each gene contains some normal samples and some cancerous samples, traditional k-means clustering algorithm [11-13] with k =2 is applied on each selected gene and miss-classification accuracy is computed based on which only the most important genes are selected for classification.
iv. Finally, classification rules [7, 14, 15] are generated for each gene on the basis of training dataset to identify cancer and non cancer samples of test dataset and obtained satisfactory accuracy.

The article is organized into four sections. Section 2 describes the proposed gene selection and classification methodology to select only the important genes according to high classification

accuracy. The experimental results and performance of the proposed method for a variety of benchmark gene expression datasets is evaluated in Section 3. Finally, conclusions are drawn in Section 4.

2. GENE SELECTION AND CLASSIFICATION

Conventionally morphological identification of cancer is not always effective as revealed by frequent occurrences of misdiagnoses. Recent molecular biological studies have concerned that cancer was a disease involving dynamic changes in the genome. Moreover, the rapid advances in cancer diagnosis technology have made it possible to simultaneously measure the expression levels of genes of microarray data in a single experiment. This technology has much facilitated the detection of cancerous molecular markers with respect to specified microarray dataset [1].
One current difficulty in interpreting microarray data comes from their innate nature of 'high dimensional large sample size'. Therefore, robust and accurate gene selection methods are required to identify differentially expressed group of genes across different samples, e.g. between cancerous and normal cells. Gene selection is necessary to find out genes, responsible for complex disease which take part in disease network and provide information about disease related genes. Successful gene selection will help to classify different cancer types, lead to a better understanding of genetic signatures in cancers and improve treatment strategies. Although gene selection and cancer classification are two closely related problems, most existing approaches handle them separately by selecting genes prior to classification.

2.1. Relevance Analysis of Genes

Let the labeled microarray gene expression dataset M$DS = (U, C, D)$, where $U = \{g_1, g_2, \ldots, g_n\}$ is the universe of discourse contained all the genes of the dataset, $C = \{C_1, C_2, \ldots, C_m\}$ is C is the condition attribute set contains all the samples and $D = \{d_1, d_2\}$ is the set of decision attributes. The Table1 shows the example of MDS with gene expression values and decision attributes.

Table1. Microarray dataset decision table (genes/samples).

		Condition attributes (Samples)						
		Decision attributes (classes)						
		Class1(d_1)			Class2(d_2)			
		S_1	S_2	….	S_i	S_{i+1}	…..	S_m
Set of Genes	g_1	M(1,1)	M(1,2)	….	M(1,i)	M(1,i+1)	…..	M(1,m)
	g_2	M(2,1)	M(2,2)	….	M(2,i)	M(2,i+1)	…..	M(2,m)
	….	…..	…..	….	…..	….	…..	…..
	g_n	M(n,1)	M(n,2)	…..	M(n,i)	M(n,i+1)	…..	M(n,m)

As all genes are not important to identification of particular cancer diseases, a relevance analysis of genes is necessary to select only the important genes. Initially, gene dataset MDS are preprocessed by standardizing the samples to z-score using Transitional State Discrimination

method (TSD) [10]. In TSD, discretization factor f_{ij} is computed for sample $C_j \in C$ of gene $g_i \in U$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, using (1).

$$f_{ij} = \frac{M_i[C_j] - \mu_i}{\delta_i} \tag{1}$$

Where, μ_i and δ_i are the mean and standard deviation of gene g_i and $M_i[C_j]$ is the value of sample C_j in gene g_i. Then mean (N_i) of negative values and mean (P_i) of positive values are computed from f_{ij} of each gene g_i and discretized to one of fuzzy linguistic term [16] and discretized to one of fuzzy linguistic term using (2).

$$f_{ij} = \begin{cases} 'VL' & if\ f_{ij} \le N_i \\ 'L' & if\ N_i < f_{ij} < 0 \\ 'Z' & if\ f_{ij} = 0 \\ 'H' & if\ 0 < f_{ij} < P_i \\ 'VH' & if\ f_{ij} \ge P_i \end{cases} \tag{2}$$

As the samples of genes are collected from both normal and cancerous patients, so the samples are divided into two disjoint classes say, d_1 and d_2. Now for each gene, frequencies of discrete sample values are computed in each class. Now for each gene i, maximum frequencies of discrete sample values are computed in each class using (3) and (4), respectively.

$$P_{li} = Count\ (\ f_{ij} | \ j = 1,2, \ldots, d_1\ and\ f_{ij} \in \{'VL', 'L', 'Z', 'H', 'VH'\}) \tag{3}$$

$$P_{ri} = Count(\ f_{ij} | \ j = 1,2, \ldots, d_2\ and\ f_{ij} \in \{'VL', 'L', 'Z', 'H', 'VH'\}) \tag{4}$$

Where, Count(x) is the numeric counting amount of maximum frequencies in class d_1 and d_2 for gene g_i respectively. If the maximum frequencies of P_{li} and P_{ri} occur for same discrete value, then the gene g_i is not so important as both the normal and cancerous samples are almost similar. Otherwise, the sample values of normal and cancerous samples are distinct for gene g_i and so the gene is considered as an important gene with importance factor (PF_i) computed using (5).

$$PF_i = \frac{P_{li} + P_{ri}}{m} \tag{5}$$

Where, $i = 1, 2, \ldots, n$ and m is the total number of samples. So, higher the importance factor more relevant the gene is and vice versa.

2.2. Reduct Generation

The measurement of similarity/dissimilarity among the genes based on the distance metric may not be effective for gene data analysis in a high dimensional space. And at the same time, elegant gene selection decreases the workload and simplifies the subsequent design process to a great extent. So, the method proposed a design approach to compute a minimum subset of genes called reduct which can, by itself, fully characterize the knowledge in the gene database as the whole set of genes (U) and preserves partition of data with respect to cancer classification. After computing importance factor of all genes, top n_1 (where, $n_1 << n$) number of genes are selected as initial reduct IRED. But in most of the cases, the initial reduct could not classify normal and cancerous

samples with high classification accuracy. As a result, some most important genes are selected from initial reduct and form final reduct FRED.

To obtain the final reduct, genes in IRED are partitioned from high dimensional space into lower dimensional space i.e., n_1 numbers of one-dimensional matrices are formed, one for each gene. Since, each gene contains some normal and some cancerous samples, it is expected that the sample values will form two disjoint clusters, one containing normal sample values and other with cancerous sample values. So traditional k-means clustering algorithm [11-13] with k =2 is applied on the gene and miss-classification accuracy is computed using (6).

$$ME_i = \frac{m_{1i}+m_{2i}}{m} \quad (6)$$

Where, m_{1i} is the number of d_1 class samples clustered as d_2 class samples and m_{2i} is the number of d_2 class samples clustered as d_1 class samples and m is the total number of samples.

In single dimensional space, k-means algorithm is very effective with respect to distance metric and also the algorithm is effective here because of limited number of genes in IRED. Final reduct FRED is formed by n_2 (where, $n_2 << n_1$) number of genes with lowest miss-classification accuracy.

Algorithm: Reduct Generation

Input: Discretized gene dataset $U = \{g_1, g_2, \ldots, g_n\}$ with sample set $C = \{C_1, C_2, \ldots, C_m\}$

Output: FRED contains most important genes.

Begin

d_1 = class in which normal samples of the genes lie

d_2 = class in which cancerous samples of the genes lie

For i=1 to n do {
P_{ki}=maximum frequency among all discrete values in d_1 of gene g_i

P_{li}=maximum frequency among all discrete values in d_2 of gene g_i

If ($P_{ki} \neq P_{li}$) then Compute importance factor PF_i of gene g_i using (5)
}

Arrange all genes in non increasing order of PF_i

IRED = set of first n_1 genes, where, $n_1 << n$

For i=1 to n_1 do {

Apply k-means clustering algorithm with k=2 on gene g_i in IRED

m_1 = number of d_1 class samples misplaced in d_2 class

m_2 = number of d_2 class samples misplaced in d_1 class

Compute mis-classification accuracy ME_i of gene g_i using (6)

}

Arrange ME_i in non decreasing order of ME_i

FRED = set of first n_2 genes, where, $n_2 << n_1$

End

2.3. Classifier Construction

The classifier is an important tool [7, 14, 15] constructed from the nature (i.e., expression values) of selected important gene of training experimental dataset for classification of cancerous and non-cancerous test samples. Here, only a set of most important genes are selected from the gene dataset and kept in FRED and classification rules are generated individually for each of the genes. Classification rules generated are of the form of "*x -> y*" indicates that "*if x, then y*", where x is the description on condition attributes or samples and y is the description on decision attributes or types of a gene. Gene is described by the sample values, some from normal and some from cancerous patients. So, two classes say, d_1 and d_2 are associated to each gene, where some sample values corresponding to d_1 and some to d_2. Let, the intervals in which the sample values of class d_1 and class d_2 are $[min_1, max_1]$ and $[min_2, max_2]$ respectively. Then one of the three different possibilities (i) non-overlapping intervals (ii) overlapping intervals and (ii) one interval fully contained in other may occurs. The rules generated in three cases are described separately.

(i) Non-overlapping intervals: Without loss of generality, assume that $max_1 < min_2$, otherwise two classes are interchanged before rule generation. Hence, gap between two intervals i.e. $(min_2 - max_1)$ is equally divided and intervals are extended accordingly. Thus the mid-point value R of the gap is considered as the upper limit of the sample values of normal genes beyond which samples are of cancerous genes, as shown in Fig. 1. So the rules are:

If (min1 <= sample value < R) then normal samples

If (R <= sample value <=max2) then cancerous samples

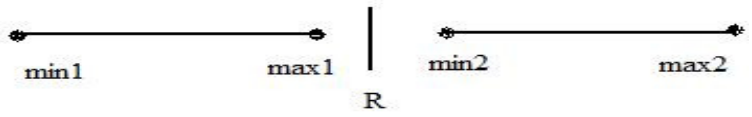


Figure1. Range of values of samples in non-overlapping intervals

(ii) Overlapping intervals: In the case, one interval is not considered as a proper subset of the other, which is described in next case. Here, also without loss of generality, assume that, $min_2 < max_1$. So, the range of overlap portion is $max_1 - min_2$. The range is not divided equally in this case, rather it is divided based on the number of samples of each class lies in it. If the ratio of percentage of samples of class d_1 to that of class d_2 in the range is m: n, then the value (R) of the point at which the range divided is obtained by (7) or (8) and R is considered as the upper limit of the sample values of normal genes beyond which samples are of cancerous genes as shown in Fig.2.

$$R = min_2 + \frac{m}{m+n} \times (max_1 - min_2) \quad (7)$$

$$R = max_1 - \frac{n}{m+n} \times (max_1 - min_2) \quad (8)$$

So the rules are:

If (min1 <= sample value < R) then normal samples

If (R <= sample value <=max2) then cancerous samples

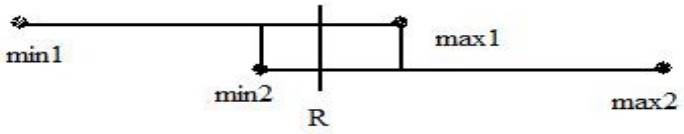


Figure2. Range of values of samples in overlapping intervals

(iii) One interval fully contained in other: Without loss of generality, assume that, class d_2 is fully contained in class d_1 such that $min_1 < min_2 < max_2 < max_1$. Here, the range ($max_2$ - min_2) contains all samples of class d_2 together with some samples of class d_1. Similar to step (ii) if the ratio of percentage of samples of class d_1 to that of class d_2 in the range is m: n, then the value (R) of the point at which the range (max_2 - min_2) divided, as shown in Fig. 3, is obtained by (9) or (10).

$$R = min_2 + \frac{m}{m+n} \times (max_2 - min_2) \quad (9)$$

$$R = max_2 - \frac{m}{m+n} \times (max_2 - min_2) \quad (10)$$

Since, class d_2 is fully contained in class d_1, the value of R may be the upper limit or lower limit of the sample values of class d2 (i.e., cancerous genes) and thus two possible rules are

(i) If (min1 <= sample value < R) OR (max2 < sample value <= max1)) then normal samples

(ii)
If (R <= sample value <=max2) then cancerous samples OR

(iii) If (min1 <= sample value < min2) OR (R < sample value <= max1)) then normal samples

If (min2 <= sample value <=R) then cancerous samples

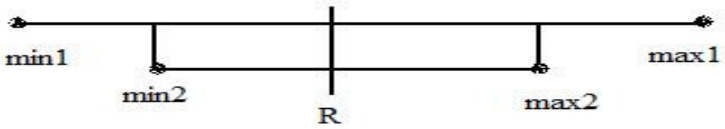


Figure3. Range of values of samples one contained in other interval

Algorithm: Classification Rule Generation

Input: Final reduct FRED with G numbers of genes and all samples of training dataset.

Output: Suitable classification rule to classify test-dataset.

Begin

For each gene g from FRED do {

d_1 = normal class associated to gene g

d_2 = cancerous class associated to gene g

Interval of sample values in d_1= [min_1, max_1] and d_2= [min_2, max_2]

Case 1:

If ($max_1 < min_2$) then {

R= max_1 + (min_2- max_1) / 2

(min1 <= sample value < R) = > d_1 (normal samples)

(R <= sample value <=max2) = > d_2 (cancerous samples)

} /*otherwise interchange d_1 by d_2 and get rules*/

Case 2:

If ($min_2 < max_1$) then {

m: n = ratio of percentage of samples in d_1 to d_2 in (max_1 - min_2)

Compute R using (7) or (8)

(min1 <= sample value < R) = > d_1 (normal samples)

(R <= sample value <=max2) = > d_2 (cancerous samples)

} /*otherwise interchange d_1 by d_2 and get rules*/

Case 3:

If ($min_1 < min_2 < max_2 < max_1$) then {

m: n = ratio of percentage of samples in d_1 to d_2 in (max_2 - min_2)

Compute R using (9) or (10)

Two possible rules are:

(i) (min1 <= sample value < R) || (max2 < sample value <= max1) => d_1 (normal samples) and (R <= sample value <=max2) => d_2 (cancerous samples)

OR

(ii) (min1 <= sample value < min2) || (R < sample value <= max1) => d_1 (normal samples) and (min2 <= sample value <=R) => d_2 (cancerous samples)

} /*otherwise interchange d_1 by d_2 and get rules*/

End

3. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

Experimental studies presented here provide an evidence of effectiveness of proposed gene selection and classification technique. Experiments were carried out on large number of different kinds of microarray data, few of them publicly available [17-21] as training and test dataset are summarized in Table 2. Each dataset contains two types of samples, one group is normal and other is cancerous.

Table2. Summary of Gene expression (training/testing) dataset.

Dataset	No.of Genes	Class Name	No. of Training Samples (class1/class2)	No.of Test Samples (class1/class2)
Leukemia	7129	ALL/AML	38(27/11)	34(20/14)
Lung Cancer	12533	MPM/ADCA	32(16/16)	149(15/134)
Prostate Cancer	12600	Tumor/Normal	102(52/50)	34(25/9)
Breast Cancer	24481	Relapse/Non-relapse	78(34/44)	19(12/7)

In addition, because there are microarray intensity discrepancies between the training set and the test set in the prostate cancer dataset [19, 20] caused by two different experiments, so normalization is required for both the training and the test dataset. Each original expression level M(i,j) is normalized using (11).

$$M(i,j)_{i=1,..,n \text{ and } j=1,..,m} = \frac{M(i,j) - \left[max_{j=1,..,m}\{M(i,j)\} + min_{j=1,..,m}\{M(i,j)\}\right]/2}{\left[max_{j=1,..,m}\{M(i,j)\} - min_{j=1,..,m}\{M(i,j)\}\right]/2} \quad (11)$$

After the normalization, all the gene expression levels are limited in interval [-1, 1]. For the other datasets, to avoid unnecessary loss of information, the normalization process is not conducted since the training and the test sets are from the same experiments [17, 18, 21].

The proposed method, computes firstly initial reduct set IRED of seventy five genes with top probability factors and then final reduct set FRED with fifteen genes with less miss-classification errors. It is observed that all final identified genes of all gene dataset are most important with respect to classification accuracy.

In Leukemia dataset [17], seven genes with their computed importance factor, mis-classification error and classification accuracy are listed in Table 3 and all other selected genes have the classification accuracy more than 73% (not shown). Two classification rules induced from training dataset by gene index 2288 are: if M(#2288) $\geq$ 929.5, then AML and if M(#2288) $<$ 929.5, then ALL. Likewise, gene #760 induces two rules: if M (Gene_id_760) $\geq$ 720.5, then AML and if M (Gene_id_760) $<$ 720.5, then ALL.

Table 3: Most important Leukemia (ALL/AML) genes

Gene_id	Gene name	Correctly classified samples [Total(ALL/AML)]	Classification accuracy (%) [Total(ALL/AML)]	Kappa Statistics	Importance Factor	Miss-classification error
2288	M84526_at	34 (21/13)	97.89 (100/93)	0.9459	0.921053	0.131579
1882	M27891_at	33 (20/13)	95.12 (96/93)	0.9078	0.894737	0.131579
1834	M23197_at	33 (19/14)	95.08 (92/97)	0.8954	0.921053	0.131579
4847	X95735_at	32 (19/13)	92.67 (91/93)	0.8650	0.973684	0.078947
760	D88422_at	32 (21/11)	91.78 (100/79)	0.8641	0.894737	0.236842
4373	X62320_at	31 (20/11)	89 (96/79)	0.8139	0.868421	0.236842
3320	U50136_rna1_at	26 (19/7)	75 (91/50)	0.7321	0.921053	0.052632

Similarly, for Lung cancer dataset [18], similar information are shown in Table 4 for fourteen genes and all other selected genes have the classification accuracy more than 80% (not shown). Two classification rules induced from training dataset by gene index 5301 are: if M (#5301) $\leq$-138.9, then MPM and if M (#5301) >-138.9 then ADCA. Likewise, gene index 7765 induces two rules: if M (Gene_id_7765) > 185.9, then MPM and if M (Gene_id_7765) $\leq$ 185.9, then ADCA.

Table 4. Most important Lung cancer (MPM/ADCA) genes.

Gene_ id	Gene name	Correctly classified samples [Total(MPM/ ADCA)]	Classification accuracy (%) [[Total(MPM/ADC A)]	Kappa Statistics	Impor- tance Factor	Miss- classifia -tion error
5301	35276 _at	145(14/131)	97.32(93.34/97.76)	0.860	0.90625	0.125
7765	37716 _at	145(11/134)	97.32(73.34/100)	0.860	0.90625	0.125
12114	575_s _at	143(14/129)	95.98(93.34/87.32)	0.7190	0.90625	0.125
8537	38482 _at	141(15/126)	94.64(100/94.03)	0.6994	0.9375	0.0625
11015	40936 _at	139(13/126)	93.29(86.67/94.03)	0.5796	0.90625	0.125
3844	33833 _at	139(13/126)	93.29(86.67/94.03)	0.5796	0.875	0.21875
3333	33327 _at	138(14/124)	92.62(93.34/92.54)	0.5493	0.9375	0.125
7249	37205 _at	134(12/122)	89.94(80/91.05)	0.4963	0.90625	0.03125
2039	32046 _at	134(12/122)	89.94(80/91.05)	0.4963	0.96875	0.03125
9863	39795 _at	133(14/119)	89.27(93.34/88.81)	0.4954	0.9375	0
11841	41755 _at	132(10/122)	88.59(66.67/91.05)	0.4851	0.90625	0.09375
9474	39409 _at	131(14/117)	87.92(93.34/87.32)	0.4822	0.96875	0.15625
3508	32046 _at	125(14/111)	83.90(93.34/82.84)	0.4377	0.96875	0.0625
1136	2047_ s_at	122(11/111)	81.88(73.34/82.84)	0.4345	0.9375	0.03125

Similarly, for Prostate cancer dataset [19, 20], similar information are shown in Table 5 for seven genes and all other selected genes have the classification accuracy more than 75% (not shown). Two classification rules induced from training dataset by gene index 6185 are: if M (#6185) > -0.716381, then Tumor and if M (#6185) ≤ -0.716381, then Normal. Likewise, gene index 3794 induces two rules: if M (#3794) ≤ -0.323077, then Tumor and if M (#3794) > -0.323077, then Normal.

Table 5. Most important Prostate cancer (Tumor/Normal) genes

Gene_ id	Gene name	Correctly classified samples [Total (Tumor/No rmal)]	Classification accuracy (%) [Total (Tumor/Normal)]	Kappa Statistics	Importance Factor	Miss- classifica- tion error
6185	37639_ at	33(24/9)	97.06(96/100)	92.80	0.852941	0.215686

3794	39939_at	32(23/9)	94.12(92/100)	0.8489	0.803922	0.215686
7557	32243_g_at	31(22/9)	91.18(88/100)	0.7982	0.794118	0.323529
10138	41288_at	31(22/9)	91.18(88/100)	0.7982	0.794118	0.235294
5757	36491_at	30(23/7)	88.24(92/77.78)	0.6756	0.754902	0.215686
9050	38044_at	29(21/8)	85.30(84/88.89)	0.6643	0.794118	0.215686
205	31444_s_at	28(19/9)	82.36(76/100)	0.6621	0.794118	0.186275

Similarly, for Breast cancer dataset [21], similar information are shown in Table 6 for seven genes and all other selected genes have the classification accuracy more than 75% (not shown). Two classification rules induced from training dataset by gene index 1505 are: if M (#1505) $\leq$ -0.005, then Relapse and if M (#1505) > -0.005, then Non-relapse. Likewise, gene index 6214 induces two rules: if M (#6214) $\leq$ -0.128, then Relapse and if M (#6214) > -0.128, then Non-relapse.

Table 6. Most important Breast cancer (Relapse/Non-relapse) genes.

Gene_id	Gene name	Correctly classified samples [Total(Relapse/Non-relapse)]	Classification accuracy (%) [Total(Relapse/Non-relapse)]	Kappa Statistics	Importance Factor	Miss-classification error
1505	AF_148505	16(10/6)	84.22(83.34/85.72)	0.8034	0.717949	0.294872
6214	NM_012429	15(10/5)	78.95(83.34/71.43)	0.7566	0.717949	0.282051
10643	NM_020974	15(9/6)	78.95(75/85.72)	0.7566	0.717949	0.307692
4732	AF_052087	15(8/7)	78.95(66.67/100)	0.7843	0.705128	0.294872
14991	Contig48590_RC	14(9/5)	73.69(75/71.43)	0.6578	0.717949	0.294872
1603	Contig46421_RC	14(10/4)	73.69(83.34/57.15)	0.6487	0.717949	0.282051
719	NM_001685	14(7/7)	73.69(53/100)	0.6732	0.74359	0.282051

The rules generated for selected genes shown in Table 3, Table 4, Table 5 and Table 6 by the proposed classification method and other methods such as Bayes classifier (Naïve Bayes), Tree based classifier (J48-C 0.25 and RandomForest), Rule based classifier (PART), Meta classifier (AdaBoostM1) and Lazy classifier (Kstar) are applied on test samples and accuracies are measured, as shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7. It is observed that for all test-dataset, the proposed and other classifiers shows better accuracy that shows the importance of selected genes. Also in most of the cases, accuracy obtained by the proposed method is higher compare to other methods which show the goodness of the proposed classifier.

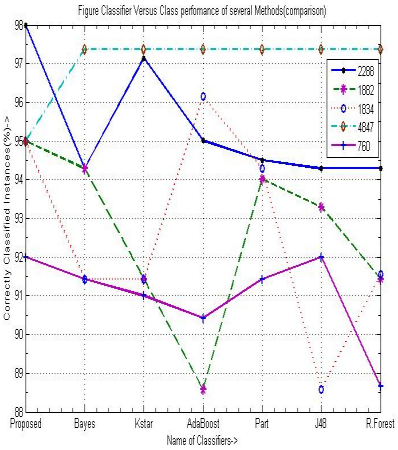


Figure 4. Performance of Leukemia genes

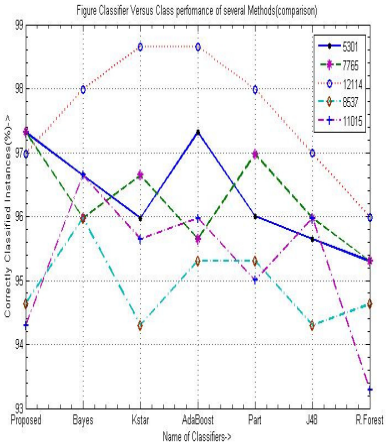


Figure 5. Performance of Lung Cancer genes

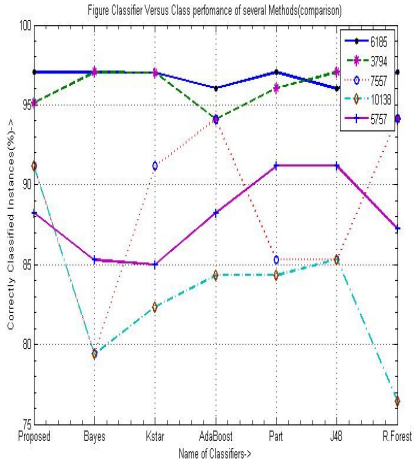


Figure5. Performance of Prostate Cancer genes

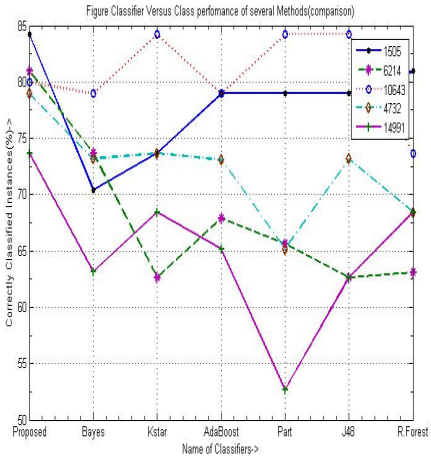


Figure 6. Performance of Breast Cancer genes

The discretization and labeling of experimental dataset are implemented using Mat lab 7.8.1 version. Also, proposed 'Reduct Generation' and 'Classification Accuracy Computation' are implemented using Mat lab 7.8.1 version and all classification performances are measured by Weaka-3-6-5 Data Mining tool [22] and comparison figures are drawn in Mat lab 7.8.1 version. The comparison is performed on PC (Intel(R) Core(TM) 2 Duo T5750 2.0 GHz, 2.0 GHz with 2.0 GB of Ram).

4. DISCUSSIONS AND CONCLUSIONS

Systematic and unbiased approach to cancer classification is of great importance to cancer treatment and drug discovery. It has been known that gene expression contains the keys to the fundamental problems of cancer diagnosis, cancer treatment and drug discovery. The recent advent of microarray technology has made the production of large amount of gene expression data possible. This has motivated the researchers in proposing different cancer classification algorithms using gene expression data.

In the paper, a novel gene selection and classification technique has been proposed for select important genes (single) and then constructs classification rules to classify cancerous and non-cancerous samples with high classification accuracy. The proposed method is applied on four publicly available experimental microarray cancer dataset and selects some important genes by comparing probability factors of all genes and form initial reduct according to proposed algorithm. Then traditional k-means algorithm is applied on initial reduct for each gene and form final reduct with more important genes on consideration of less miss-classification accuracy. Then construct classification rules on the basis of selected genes (single train gene) and classification accuracy in terms of correctly classified instances apply on test genes that shows quantitative satisfactory results. Gene selection, an important preprocessing step was presented in detail and evaluated for their relevance in cancer classification. Comparative study is also made with respect to correctly classified instances (%) by some traditional classifiers namely Bayes, J48, PART, MLP, Random Forest, AdaBoost and Kstar which shows that the goodness of the proposed method.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2 Aerman D.A., Gish K., Ybarra S., Mack D., & Levine A.J. .,(1999) "Expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proc. Natl. Acad. Sci, vol 1, pp 6745–6750.

[3] DeRisi J, et al. (1996) "Use of a cDNA microarray to analyse gene expression patterns in human cancer", Nat Genet, Dec, vol. 14, No. 4, pp 457-60.

[4] Muralidhar K. & Sarathy R., (1999) "Security of random data perturbation methods", ACM Trans. Database Syst., Vol. 24, No. 4, pp 487–493.

[5] Petrov A. & Shams S., (2004) "Microarray image processing and quality control", VLSI Signal Processing, vol. 38, No. 3, pp 211–226.

[6] Su Y., Murali T. M., Pavlovic V., Schaffer M. & Kasif S., (2003) "RankGene: identification of diagnostic genes based on expression data", BIOINFORMATICS, vol. 19, pp. 1578-1579.

[7] Li, L., Weinberg, R. C., Darden, T. A. & Pedersen L. G., (2001) "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," BIOINFORMATICS, vol. 17, pp.1131-1142.

[8] Zhang H., Yu C. Y., Singer B. & Xiong M., (2001) "Recursive partitioning for tumor classification with gene expression microarray data," PNAS, vol. 98, pp. 6730-6735.

[9] Dudoit S., Fridlyand J., & Speed T. P., (2002) "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistical Assoc., vol. 97, No. 457, pp. 77-87.

[10] Wang, X., & Gotoh, O., (2009) "Microarray-Based Cancer Prediction Using Soft Computing Approach", Cancer Informatics, vol. 7, pp 123–139.

[11] R.G. Pensa, C. Leschi, J. Besson, & J. Boulicaut., (2004) "Assessment of discretization techniques for relevant pattern discovery from gene expression data", In 4th Workshop on Data Mining in Bioinformatics.

[12] Qu Y., & Xu S., (2004) "Supervised cluster analysis for microarray data based on multivariate Gaussian mixture", Bioinformatics, vol. 20, pp 1905-13.

[13] Guha, S., Rastogi R. & Shim K., (1998) "CURE: an efficient clustering algorithm for large databases", Proc. of ACM SIGMOD International Conference on Management of Data, pp. 73 – 84.

[14] Bradley P. S., Bennett K. P. & Demiriz A., (2000) "Constrained k-means clustering (Technical ReportMSR-TR-2000-65)", Microsoft Research, Redmond, WA.

[15] Dudoit S., Fridlyand J., & Speed T.P., (2002) "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistical Assoc., vol. 97, no. 457, pp. 77-87.

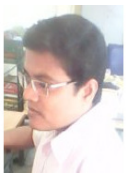
[16] Golub. T. R., (1999) "Molecular classification of cancer: class discovery and class prediction by Gene Expression Monitoring," Science, vol. 286, pp 531-537.

[17] Ivars Peterson, (1993) "Fuzzy Sets", Science News, Vol. 144, July 24, pp. 55.
[18] Leukemia dataset: http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi.
[19] Lung dataset: http://www genome. wi.mit.edu/mpr/lung.
[20] Prostate cancer train dataset: http://www-genome.wi.mit.edu/mpr/prostate.
[21] Prostate cancer test dataset: http://carrier.gnf.org/welsh/prostate.
[22] Breast cancer dataset: http://www.rii.com/publications/2002/vantveer.htm.
[23] WEKA: Machine Learning Software, http://www.cs.waikato.ac.nz/~.html

.

Authors

Mr. Soumen Kumar Pati is an Assistant Professor of Computer Science/Information Technology at St. Thomas' College of Engineering and Technology, Kidderpore, Kolkata,West Bengal, India. He has received M.Tech degree in Computer Science and Engg from Jadavpur University. He is registered for PhD (Engg) degree at Bengal Engineering and Science University, Shibpur, Howrah. His research interests include Bio-informatics, Data Mining and Pattern Recognition, Rough Set Theory, etc.



Dr.Asit Kr. Das is an Assistant Professor of Computer Science and Technology at Bengal Engineering and Science University, Shibpur, Howrah. He has received B.Sc. Honours in Mathematics, B. Tech. and M.Tech degree in Computer Science and Engg from Calcutta University. He obtained PhD (Engg) degree from Bengal Engineering and Science University, Shibpur, Howrah. His research interests include Data Mining and Pattern Recognition, Text Categorization, Rough Set Theory, Bio-informatics etc.