**A Novel Important Genes Selection and Classifier Construction for Microarray dataset**

**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 (single) 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 article, a novel method for the classification of cancer based on gene expression profiles using single gene with high class-discrimination capability according to their depended degree by the classes.

The proposed method is presented for selection of important genes and classification of cancer based on gene expression profiles using single important genes. The proposed method first computes probability 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 probability factor of each gene and form initial reduct. Then traditional k-means clustering algorithm is applied on each 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 other classifiers.

**Keywords:** Microarray cancer data, K-means algorithm, important gene selection, Decision rule, Classifier construction, Cancer classification.

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 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 [3], the 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 [4]. However, when examining microarray data, the size of the data sets and noise contained within the data sets compromises precise qualitative and quantitative analysis[5].

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 [7, 8] on the basis of the genes is another mission. If we survey the established investigations in this field, we will find that almost all the accurate classification results are obtained based on more than two genes [9, 10].

In the article, a novel gene selection technique and suitable classification rule generation technique has been proposed on projecting microarray data (high dimensional space Sh) for selecting important genes to predict cancer by using single gene. The method can be broken down into following four steps:

1. The dataset is standardized to Z-score using Transitional State Discrimination method [11] and each sample is characterized by four discrete values. After discretizing the gene expression data, the discrete value is represented by some suitable linguistic term.
2. Initial important genes are selected by calculating probability factor of each gene with respect to linguistic values of different predefined class label of microarray training-data and form initial reduct, discussed in Section 2.
3. After initial important genes are selected then traditional k-means clustering technique [12, 13] (k=2) is applied on each gene with original gene expression patterns and miss-classification errors are computed. Then some final important genes are selected according to less miss- classification errors of genes and form final reduct, discussed in Section 2.
4. Then classifier is constructed by generating suitable classification rule on the basis of training dataset to identify cancer and non cancer samples of test dataset and obtained satisfactory accuracy, discussed in Section 3.

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.

1. 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 [15] 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 [16] 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.

The objective of gene selection is two-fold: to provide a better understanding of the underlying biological system [ ] that generates data and to improve the prediction performance of classifiers. Effective gene selection often leads to a compact classifier with better accuracy and interpretability (Kitter, 1986).

* 1. **Relevance Analysis of Genes to Compute Probability Factor**

As all genes are not relevant to identification of particular cancer diseases, so a relevance analysis of genes is necessary to select only the important genes. In the proposed method, a probability factor is computed for each gene using counting of similar types of expression values for which linguistic value (according to discrete value) and decision attribute is required. So, before probability factor computation, the datasets are preprocessed by standardizing the samples to z-score using Transitional State Discrimination method (TSD) [ ]. In TSD, discretization factor *fij* is computed for sample value Cj ∈ C of gene *Mi* *U*, *i* = 1, 2, …, *n*, j = 1, 2, …, m , using (1).

Where, *μi* and *δi* are the mean and standard deviation of gene *Mi* and M*i*[*Cj*] is the value of sample *Cj* in gene M*i*. Then negative average (Ni) and positive average (Pi) are computed of for each gene *Mi* and labeled suitable linguistic value to each cell of M*i*[*Cj*]. Now the value M*i*[*Cj*] is discredited to one of ‘VL’ (very low), ‘L’ (low), ‘Z’ (zero), ‘H’ (high) and ‘VH’ (very high) depending on fij is ‘fij <=Ni ’, ‘Ni< fij<0’, ‘fij =0’, ‘0< fij< Pi’ and ‘fij >= Pi’ respectively for corresponding M*i* of M*i*[*Cj*]. Let the labeled microarray gene expression dataset M*DS* = (*U*, *C*, *D*), where *U* = {*u*1, u2, …, *u*n} is the set of genes and n is the number of genes, *C* = {*C*1, *C*2, …, *C*m} is the set of samples and m is the number of samples and *D* = {*d*1, d2} is the set of decision attributes. The Table1 shows the example of MDS with linguistic gene expression values and decision attributes.

**Table 1:** Microarray dataset decision table (genes/samples)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Samples | | | | | | | | |
| Decision attributes (classes) | | | | | | | | |
| Class1(d1) | | | | | Class2(d2) | | | |
| S1 | S2 | S3 | …. | Si | Si+1 | Si+2 | ….. | Sm |
| Condition attributes (genes) | G1 | M(1,1) | M(1,2) | M(1,3) | …. | M(1,5) | M(1,6) | M(1,7) | ….. | M(1,m) |
| G2 | M(2,1) | M(2,2) | M(2,3) | …. | M(2,5) | M(2,6) | M(2,7) | ….. | M(2,m) |
| …. | ….. | ….. | ….. | …. | ….. | …. | ….. | ….. | ….. |
| Gn | M(n,1) | M(n,2) | M(n,3) | ….. | M(n,5) | M(n,6) | M(n,7) | ….. | M(n,m) |

Now probability factors (PFi) are computed for each gene (ui) by counting distinct linguistic discrete label for individual decision attributes. Let Pk and Pl are the maximum count of any discrete label of d1 (class1) and d2 (class2) for any ui, respectively. If Pk and Pl related discrete label are different then PFi are computed for each individual gene, by using (2).

Where, *i* = 1, 2, …, *n* is the total number of genes and m is the total number of samples. If Pk and Pl related linguistic discrete labels are different with respect to d1 and d2 and is higher for a particular gene then this gene is more important, otherwise the gene is less important. Because, if the discrete label are different with respect to decision attributes then expression values are more similar for one class (d1) and dissimilar for another class (d2) for a gene. If the gene expression values vary in different classes for a gene then samples are easily classified by any classifier [ ] using some classification rules which are generated from selected important genes.

* 1. **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 probability factor for all genes, top n1 (where, n1<<n) numbers of genes are selected according to probability factors and form initial reduct IRED. But in most of the cases, the initial reduct could not classify normal and cancerous samples with high classification accuracy. So, the proposed algorithm select some most important genes from initial reduct and form final reduct FRED. To obtain the final reduct, initially, IRED dataset is partitioned from high dimensional space into lower dimensional space i.e., n1 numbers of one dimensional matrices are formed with respect to individual genes. Then apply traditional k-means clustering algorithm [ ] (where, number of clusters, k=2) on single dimensional matrices separately and miss-classification accuracy are computed by using (3).

Where, m1 is the number of class1 (d1) test-samples clustered as class2 (d2) samples, m2 is the number of class2 (d2) test-samples clustered as class1 (d1) samples and m is the total number of test samples. In single dimensional space, k-means algorithm is very effective with respect to distance metric. There are hundred or more samples are available in test-dataset and pre-defined decision numbers (i.e., d1 and d2) are also present in test-dataset. So, miss-classification accuracy is easily computed using (3) by applying k-means algorithm [ ] in one dimensional space. With respect to time complexity, k-means algorithm is also applicable, because there are limited numbers of samples (i.e., Expression values) and applied on IRED only. After computing miss-classification errors (), the are arranged in non-decreasing order and n2 (where, n2<<n1) numbers of genes are selected according to less miss-classification errors and form final reduct FRED.

**Algorithm: Reduct Generation**

**Input:** Linguistic gene expression value i.e., .

**Output:** Final important genes with importance factors and misclassification errors.

Begin

For i=1 to n do

Count maximum linguistic label of class1 samples, say L1.

Count maximum linguistic label of class2 samples, say L2.

If ( then /\* and both are distinct discrete label\*/

Compute probability factor, using (2).

End if

Arrange PFi in non increasing order and select important n1 genes according to PFi where n1<< n and form initial reduct IRED.

End for

For i=1 to n1 do

Apply k-means clustering technique for k=2 on each initial selected gene and compute miss-classification errors, , using (3).

Arrange in non increasing order and select important n2 genes according to where n2<< n1 and form final reduct FRED.

End for

End

* 1. **Classifier Construction**

Now, the classification rules are required to classify cancerous and non-cancerous samples and these are generated from the nature (i.e., expression values) of selected important gene of training experimental dataset. Inducing classification rulesfrom decision tables is one of the main tasks in proposed method. One decision rule in the form of “*x* y” indicates that “if *x*, then y”, where xis the description on condition attributes and y is the description on decision attributes. In the proposed method, the decision rules are constructed from an important gene of final reduct. The rules are generated from the following algorithm and get satisfactory results by applying on test-dataset in terms of correctly classified samples.

**Algorithm:** **Classification Rule Generation**

**Input:** Final reduct FRED with n2 numbers of genes and all samples of training dataset.

**Output:** Suitable classification rule to classify test-dataset.

Begin

Take one important gene from FRED and find out minimum and maximum expression value from class1 and class2, say min1, max1 and min2, max2 respectively.

Case1: If (max1 < min2) { /\* All class1 expression values are numerically less than all class2 expression values\*/

d= | max1- min2|

R= max1 + (d/2).

}

Case2: Else if (min1> max2) { /\* All class1 expression values are numerically greater than all class2 expression values\*/

d= | min1- max2|

R= max2+ (d/2).

}

Else { /\*some class1 expression values numerically greater than minimum expression value of class2 expression values\*/

Case3: Compute number of samples for expression values >= min2 from class1, say d1 and also compute number of samples for expression values<= max1 from class2, say d2.

Then,

R= max1+d.

Or, /\*some class1 expression values numerically less than maximum expression value of class2 expression values\*/

Case4: Compute number of samples for expression values <= max2 from class1, say d1 and also compute number of samples for expression values >= min1 from class2, say d2.

Then,

R= min1+d.

}

End if

If (expression value<=R) then class1 samples /\*classification rules\*/

Else class2 samples (for Case1 and Case3).

If (expression value>R) then class1 samples

Else class2 samples (for Case2 and Case4).

End

1. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

Experimental studies presented here provide an evidence of effectiveness of proposed gene selection and classification technique. The four microarray cancerous dataset are publicly available as training and test dataset used in the experiment. Those are described below where general description of dataset presents in Table 1.

* 1. **Datasets Description**

Experiments were carried out on large number of different kinds of microarray data (cancerous data) [ ], few of them described below are summarized. Each dataset contains two types of samples, one group is normal and other is cancerous. The gene number, class, training and test sample numbers contained in the four datasets are listed in Table 2.

**Table 2:** Summery of four Gene expression (training/testing) dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset Name | No. of Genes | Class Name | No. of Training Samples (class1/class2) | No. of Test Samples (class1/class2) | Raw Data Availability |
| Leukemia | 7129 | ALL/AML | 38(27/11) | 34(20/14) | <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi> |
| Lung Cancer | 12533 | MPM/ADCA | 32(16/16) | 149(15/134) | <http://www> genome. wi.mit.edu/mpr/lung |
| Prostate Cancer | 12600 | Tumor/Normal | 102(52/50) | 34(25/9) | (1) Train data: http://www-genome.wi.mit.edu/mpr/prostate  (2) Test data: http://carrier.gnf.org/welsh/prostate |
| Breast Cancer | 24481 | Relapse/Non-relapse | 78(34/44) | 19(12/7) | http://www.rii.com/publications/2002/vantveer.htm |

* 1. **Results and Analysis**

In the proposed method, the original training dataset are discretized by the Transitional State Discrimination method (TSD) [ ] using (1). Every continuous-valued expression is discretized into the attribute with no more than four different values discussed in Section 2. In addition, because there is microarray intensity discrepancies between the training set and the test set in the prostate cancer dataset 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 (4).

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 [ ].

In all the dataset, the initial reduct IRED contains seventy five genes with top probability factors according to Reduct Generation algorithm. Then final reduct contains fifteen genes with less miss-classification errors according proposed algorithms. And all final identified genes are most important with respect to classification accuracy that shows in different Tables.

In Leukemia dataset, six genes have the classification accuracy no less than 89% and all other genes the classification accuracy no less than 73% among final identified fifteen genes. The gene expression levels are denoted by M(i), where, i=1,2,…,m for a particular gene and two classification rules induced from training dataset by gene index 2288 are: if M(Gene\_id\_2288) ≥ 929.5, then AML and if M(Gene\_id\_2288) < 929.5, then ALL. Likewise, gene #760 induces two rules: if M (Gene\_id\_760) ≥ 720.5, then AML and if M (Gene\_id\_760) < 720.5, then ALL. Table 3 summarizes the information on the some genes of test dataset with no less than 75% classification accuracy.

**Table 3:** Genes with high classification accuracy in the Leukemia (ALL/AML) dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene\_id | Accession | Correctly classified samples [Total(ALL/AML)] | Classification accuracy (%) [Total(ALL/AML)] | Probability Factor | Miss-classification error |
| 2288 | M84526\_at | 34 (21/13) | 98 (100/93) | 0.921053 | 0.131579 |
| 1882 | M27891\_at | 33 (20/13) | 95 (96/93) | 0.894737 | 0.131579 |
| 1834 | M23197\_at | 33 (19/14) | 95 (92/97) | 0.921053 | 0.131579 |
| 4847 | X95735\_at | 32 (19/13) | 92 (91/93) | 0.973684 | 0.078947 |
| 760 | D88422\_at | 32 (21/11) | 92 (100/79) | 0.894737 | 0.236842 |
| 4373 | X62320\_at | 31 (20/11) | 89 (96/79) | 0.868421 | 0.236842 |
| 3320 | U50136\_rna1\_at | 26 (19/7) | 75 (91/50) | 0.921053 | 0.052632 |

In Lung cancer dataset, seven genes have the classification accuracy no less than 92% and all other genes the classification accuracy no less than 80% among the final identified fifteen genes. The gene expression levels are denoted by M(i), where, i=1,2,…,m for a particular gene and two classification rules induced from training dataset by gene index 5301 are: if M(Gene\_id\_5301) ≤-138.9, then MPM and if M(Gene\_id\_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) ≤ 185.9, then ADCA. Table 4 summarizes the information on the some genes of test dataset with no less than 81% classification accuracy.

**Table 4:** Genes with high classification accuracy in the Lung cancer (MPM/ADCA) dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene\_id | Accession | Correctly classified samples [Total(MPM/ADCA)] | Classification accuracy (%) [[Total(MPM/ADCA)] | Probability Factor | Miss-classification error |
| 5301 | 35276\_at | 145(14/131) | 97.32(93.34/97.76) | 0.90625 | 0.125 |
| 7765 | 37716\_at | 145(11/134) | 97.32(73.34/100) | 0.90625 | 0.125 |
| 12114 | 575\_s\_at | 143(14/129) | 95.98(93.34/87.32) | 0.90625 | 0.125 |
| 8537 | 38482\_at | 141(15/126) | 94.64(100/94.03) | 0.9375 | 0.0625 |
| 11015 | 40936\_at | 139(13/126) | 93.29(86.67/94.03) | 0.90625 | 0.125 |
| 3844 | 33833\_at | 139(13/126) | 93.29(86.67/94.03) | 0.875 | 0.21875 |
| 3333 | 33327\_at | 138(14/124) | 92.62(93.34/92.54) | 0.9375 | 0.125 |
| 7249 | 37205\_at | 134(12/122) | 89.94(80/91.05) | 0.90625 | 0.03125 |
| 2039 | 32046\_at | 134(12/122) | 89.94(80/91.05) | 0.96875 | 0.03125 |
| 9863 | 39795\_at | 133(14/119) | 89.27(93.34/88.81) | 0.9375 | 0 |
| 11841 | 41755\_at | 132(10/122) | 88.59(66.67/91.05) | 0.90625 | 0.09375 |
| 9474 | 39409\_at | 131(14/117) | 87.92(93.34/87.32) | 0.96875 | 0.15625 |
| 3508 | 32046\_at | 125(14/111) | 83.90(93.34/82.84) | 0.96875 | 0.0625 |
| 1136 | 2047\_s\_at | 122(11/111) | 81.88(73.34/82.84) | 0.9375 | 0.03125 |

In Prostate cancer dataset, five genes have the classification accuracy no less than 88% and all other genes the classification accuracy no less than 80% among the final identified fifteen genes. The gene expression levels are denoted by M(i), where, i=1,2,…,m for a particular gene and two classification rules induced from training dataset by gene index 6185 are: if M(Gene\_id\_6185) > -0.716381, then Tumor and if M(Gene\_id\_6185) ≤ -0.716381, then Normal. Likewise, gene index 3794 induces two rules: if M (Gene\_id\_3794) ≤ -0.323077, then Tumor and if M (Gene\_id\_3794) > -0.323077, then Normal. Table 5 summarizes the information on the some genes of test dataset with no less than 82% classification accuracy.

**Table 5:** Genes with high classification accuracy in the Prostate cancer (Tumor/Normal) dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene\_id | Accession | Correctly classified samples [Total (Tumor/Normal)] | Classification accuracy (%) [Total (Tumor/Normal)] | Probability Factor | Miss-classification error |
| 6185 | 37639\_at | 33(24/9) | 97.06(96/100) | 0.852941 | 0.215686 |
| 3794 | 39939\_at | 32(23/9) | 94.12(92/100) | 0.803922 | 0.215686 |
| 7557 | 32243\_g\_at | 31(22/9) | 91.18(88/100) | 0.794118 | 0.323529 |
| 10138 | 41288\_at | 31(22/9) | 91.18(88/100) | 0.794118 | 0.235294 |
| 5757 | 36491\_at | 30(23/7) | 88.24(92/77.78) | 0.754902 | 0.215686 |
| 9050 | 38044\_at | 29(21/8) | 85.30(84/88.89) | 0.794118 | 0.215686 |
| 205 | 31444\_s\_at | 28(19/9) | 82.36(76/100) | 0.794118 | 0.186275 |
| 12148 | 575\_s\_at |  |  | 0.803922 | 0.235294 |

In Breast cancer dataset, seven genes have the classification accuracy no less than 73% and all other genes the classification accuracy no less than 68% among the final identified fifteen genes. The gene expression levels are denoted by M(i), where, i=1,2,…,m for a particular gene and two classification rules induced from training dataset by gene index 1505 are: if M(Gene\_id\_1505) ≤ -0.005, then Relapse and if M(Gene\_id\_1505) > -0.005, then Non-relapse. Likewise, gene index 6214 induces two rules: if M (Gene\_id\_6214) ≤ -0.128, then Relapse and if M (Gene\_id\_6214) > -0.128, then Non-relapse. Table 6 summarizes the information on the some genes of test dataset with no less than 81% classification accuracy.

**Table 6:** Genes with high classification accuracy in Breast cancer (Relapse/Non-relapse) dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene\_id | Accession | Correctly classified samples [Total(Relapse/Non-relapse)] | Classification accuracy (%) [Total(Relapse/Non-relapse)] | Probability Factor | Miss-classification error |
| 1505 | AF148505 | 16(10/6) | 84.22(83.34/85.72) |  |  |
| 6214 | NM\_012429 | 15(10/5) | 78.95(83.34/71.43) |  |  |
| 10643 | NM\_020974 | 15(9/6) | 78.95(75/85.72) |  |  |
| 4732 | AF052087 | 15(8/7) | 78.95(66.67/100) |  |  |
| 14991 | Contig48590\_RC | 14(9/5) | 73.69(75/71.43) |  |  |
| 1603 | Contig46421\_RC | 14(10/4) | 73.69(83.34/57.15) |  |  |
| 719 | NM\_001685 | 14(7/7) | 73.69(53/100) |  |  |

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. 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).

By the help of identified important genes (single), the samples are classified by the proposed classification method and other methods [ ] such as Bayes classifier (Naïve Bayes), Trees based classifier (J48-C 0.25), Rules based classifier (PART), Trees based classifier (RandomForest), Meta classifier (AdaBoostM1), and Lazy classifier (Kstar) and accuracies are plotted with various colors for each selected gene, as shown in Fig. 1 to Fig. 4. It is observed that for all test-dataset, the proposed and other classifiers shows better accuracy that shows the impotency of identified genes. And most of cases, the comparison figures shows that the goodness of proposed classifier with comparison to other classifiers according to the better classification results (%). All classification performances are measured by Weaka-3-6-5 Data Mining tool and comparison figures are drawn in Mat lab 7.8.1 version.



Fig. 1: Classification Performance of selected genes for Leukemia dataset



Fig. 2: Classification Performance of selected genes for Lung Cancer dataset



Fig. 3: Classification Performance of selected genes for Prostate Cancer dataset



Fig. 4: Classification Performance of selected genes for Breast Cancer dataset

1. DISCUSSION AND CONCLUSION

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 are computed apply on test gene that shows quantitative satisfactory results. In the paper, the impotencies of selected genes are evaluated by their performance in three aspects: computation time, classification accuracy and statistical significance. Gene selection as an important preprocessing step was also 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.