Mutual Information and Cross Entropy Framework to Determine Relevant Gene Subset for Cancer Classification

Rajni Bala
Deen Dayal Upadhyaya College, University of Delhi, Delhi, India
E-mail: rbala@ddu.du.ac.in

R.K. Agrawal
School of Computer and Systems Sciences, Jawaharlal Nehru University, Delhi, India
E-mail: rka@mail.jnu.ac.in

Keywords: cancer classification, microarray datasets, mutual information, cross entropy, gene selection

Received: August 5, 2010

Classification of microarray datasets has drawn attention of research community in last few years. Microarray datasets are characterized by high dimension and small sample size. To avoid curse of dimensionality good gene selection methods are needed. Here, we propose a two stage algorithm MICE for finding a small subset of relevant genes responsible for classification of high dimensional microarray datasets. The proposed method is based on the principle of Mutual Information and Cross 'Entropy. In first stage of algorithm, mutual information is employed to select a set of relevant genes and cross entropy is used to determine independent genes. In second stage, a wrapper based forward feature selection method is used to obtain a set of optimal genes for a given classifier. The efficacy of proposed algorithm is tested on seven well known publicly available microarray datasets. Comparison with other state-of-art methods shows that our proposed algorithm is able to achieve better classification accuracy with less number of genes.

Povzetek: Opisana je metoda za določanje relevantnih skupin genov za rakave bolezni.

1 Introduction

In last few years, classification of microarray datasets has drawn attention of research community. Various machine learning and data mining methods have been applied for classification of microarray datasets. But classification of microarray datasets faces many challenges. One of the main challenges is that such datasets are characterized by large number of genes and small number of samples. This small number of samples compared to the large number of genes wakes up the curse of dimensionality [2]. Also, many of these genes are not relevant to discriminate samples. These irrelevant genes not only have negative effect on the classification accuracy of the classifier but also increase data acquisition cost and learning time. For better classification there is a need to reduce dimension of such datasets.

Dimension Reduction can be done in two ways: feature selection and feature extraction [8]. Feature Selection refers to reducing the dimensionality of measurement space by discarding redundant, noisy and irrelevant features. It leads to saving in measurement cost and the selected features retain their original physical interpretation. In addition, the retained features may be important for understanding the physical process that generates patterns. Feature extraction methods like Principle Component Analysis, Independent Component Analysis utilize all the information contained in the measurement space to obtain a new transformed space and then important features are selected from the new transformed space. Transformed features generated by feature extraction methods may provide a better discriminative ability than the best subset of given features, but these new features may not provide any physical meaning. The choice between feature selection and feature extraction depend on the application domain and specific training data available. In microarray datasets one is not only interested in classifying the sample based on gene expression but also in identifying important genes/features. Hence dimension reduction is normally carried out with feature selection rather than feature extraction. Therefore, efficient feature/gene selection methods are necessary for selecting a small set of informative features/genes. Gene selection not only allows for faster and efficient model building by removing irrelevant, redundant and noisy features but also provides better understanding of genes which lead to a particular disease.

There are numbers of feature selection methods proposed in last few years. These methods broadly fall into two categories: filter and wrapper methods [8]. Most filter methods employ statistical characteristics of data for feature selection which needs less computation. They independently measure the importance of features without involving any learning algorithm. The filter approach does not take into account the learning bias introduced by the final learning algorithm, so it may not be able to select the most relevant set of features for the learning algorithm. Wrapper methods use learning algorithm for selecting feature set. It tends to find features better suited to the predetermined learning algorithm resulting in better performance. But, it is computationally more expensive since the classifier must be trained for each candidate subset. Hence, when number of features is large wrapper approaches become unfeasible.

Many filter based feature/gene selection methods have been proposed in literature [1, 6, 14, 17, 22, 23]. Broadly they are categorized in two categories: univariate and multivariate evaluation methods. Univariate evaluation methods evaluate the relevance of each feature individually. They are simple and fast therefore appealing and popular [3, 7, 13, 26]. However, they assume that the features are independent of each other. Multivariate approaches, on the contrary, evaluate the relevance of features considering how they function as a group, taking into consideration their dependency [4, 5, 9].

One of the univariate methods determines the relevance of genes by computing the mutual information between each gene and the class label. The genes are ranked based on their relevance with class i.e. in decreasing order of their mutual information and then top *m* genes are selected. In literature, it has been observed that the combination of individual good genes does not necessarily lead to good classification performance. Also, since genes are selected based on the correlation between individual gene and target class, it doesn't capture the correlation among genes. Hence gene subset so obtained may contain redundant genes. A good gene selection method is the one that not only selects the relevant genes but also reduces redundancy among the selected gene subset. In literature, some multivariate methods have been suggested to reduce the redundancy among the selected set of genes [1, 15, 22]. However, these methods consider weighted average of only pairwise correlation instead of considering joint correlation among a set of features. Hence these approaches may not select the optimal feature subset in the presence of large number of redundant features.

In this paper, we have proposed a two stage algorithm MICE for determining an optimal feature subset. In the first stage, a pool of relevant and independent genes is created using Mutual Information (MI) and cross-entropy (CE). In second stage, forward feature selection is used to find a compact feature subset that minimizes classification error (maximizes classification accuracy), from the candidate feature set.

This paper is organized as follows – Section 2 presents some of feature selection methods based on mutual information. Section 3 includes proposed algorithm MICE for gene selection based on mutual information and cross entropy. Experimental results on some well-known publicly available datasets are presented in Section 4. Conclusions are drawn in Section 5.

2 Mutual Information

Mutual information (MI) measures the dependency between a feature set and target class. Mutual information $I(\mathbf{X}_m;\mathbf{C})$ between set of *m* features ($\mathbf{X}_m$) and class label ($\mathbf{C}$) is given by

$$I(\mathbf{X}_m;\mathbf{C}) = \int\int p(\mathbf{X}_m,\mathbf{C}) \log \frac{p(\mathbf{X}_m,\mathbf{C})}{p(\mathbf{X}_m)f(\mathbf{C})} d\mathbf{X}_m d\mathbf{C} \quad (1)$$

Higher value of MI means given feature set represents the class well. Battiti [1] defined feature selection problem as the process of selecting the most relevant *m* features from the initial set of *d* features. Ideally, problem can be solved by maximizing $I(\mathbf{X}_m;\mathbf{C})$, the joint entropy between set of features $\mathbf{X}_m$ and the target class $\mathbf{C}$. But it is often difficult to estimate the joint probability density $p(\mathbf{X}_m)$. For this a greedy feature selection method based on mutual information (MIFS) was proposed by Battiti. Battiti [1] adopted a heuristic criterion for approximating the ideal solution. Battiti's MIFS selects the subset of features which maximizes the information about the class corrected by subtracting a quantity proportional to average MI with the previously selected features. Kwak and Choi [15] proposed a greedy feature selection method called MIFS-U which provides a better estimate of the MI between input features and target class in comparison to MIFS. Peng et al. [22] suggested another variant of Battiti's MIFS as min-redundancy and max-relevance (mRMR) criterion. In this work a heuristic framework was suggested to minimize redundancy and maximize relevance to select important features incrementally [22]. In this incremental approach, m^th feature x_j can be selected from the remaining set of features which maximizes the following criterion:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, C) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right] \quad (2)$$

where S_{m-1} is the set of already selected (m-1) features.

However, MIFS, MIFS-U and mRMR algorithms use incremental search approach which considers weighted average of only pair-wise correlation instead of considering joint correlation among a set of features. Hence, these approaches might not select the optimal feature subset in presence of large number of redundant features. In this paper, we have calculated the joint MI between a set of features and target class C under the assumption that data follows multivariate normal distribution. We have employed cross entropy to determine dependency among selected genes. By combining MI and CE measures we are able to obtain a reduced set of non-redundant relevant genes. This allows us to use a wrapper based forward feature selection at the second stage to search a compact feature subset, from the above gene set, that maximizes classification accuracy.

3 Proposed Method

To measure the relevance of a gene subset, mutual information between gene subset and target class is calculated for which we should have the knowledge of the joint probability density. In reality, the probability density is not known. So, we can assume parametric form of p($\mathbf{X}_m$) and p($\mathbf{X}_m$|$\mathbf{C}$) and the parameters involved in parametric form of probability density can be estimated from the observed data. Here, we assume that the probability density p(X_m|C*)* follows multivariate normal density which is given by

$$p(X_m \mid C) = \frac{1}{(2\pi)^{\frac{m}{2}} \mid \Sigma_c \mid^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(X_m - \mu_c)^t (\Sigma_c)^{-1} (X_m - \mu_c)\right] \qquad (3)$$

where μ_c, Σ_c are respectively mean and covariance of class C data. p($\mathbf{X}_m$) is also approximated by a multivariate normal density with mean μ and covariance Σ. Under this approximation, the closed form expression for MI is known. According to Padmanabhan and Dharanipragada [21], for multivariate normal density, the upper bound on mutual information is given by

$$I_{upp}(X_m ; C) = \frac{1}{2} \log \mid \Sigma \mid - \frac{1}{2} \sum_c p_c \log \mid \Sigma_c \mid \qquad (4)$$

Equ. (4) gives an upper bound on joint mutual information between m-dimensional gene vector $\mathbf{X}_m$ and the target class **C**. However, the selected gene subset $\mathbf{X}_m$ may contain redundant features which can degrade the performance of the classifier. We can reduce the redundancy by using cross-entropy. The cross entropy D [12], measures the difference between the two probability distributions $f(X)$ *and* $q(X)$ and is given by

$$D\big(f(X), q(X)\big) = \int f(X) \log \frac{f(X)}{q(X)} dX \qquad (5)$$

If we consider

$$f(X) = p(x_1, x_2, \ldots, x_m) \qquad (6)$$

And $q(X) = p(x_1)p(x_2)\ldots p(x_m)$ (7)

Then D takes the following form

$$D_m = \int \ldots . \int p(x_1, x_2, x_3, \ldots, x_m) \log \frac{p(x_1, x_2, x_3, \ldots, x_m)}{p_1(x_1)p_2(x_2)\ldots p_m(x_m)} \qquad (8)$$
$$dx_1 dx_2 \ldots . dx_m$$

If $x_1, x_2, x_3, \ldots, x_m$ are statistically independent, then $p(x_1, x_2, \ldots, x_m) = p_1(x_1)p_2(x_2) \ldots .. p_m(x_m)$. In this case D_m becomes zero. When features are not statistically independent, D_m is given by

$$D_m = \int \ldots . \int p(x_1, x_2, \ldots, x_m) \log[p(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots . dx_m \qquad (9)$$
$$- \sum_{i=1}^{m} \int p_i(x_i) \log p_i(x_i) dx_i$$

$$D_m = -S + \sum_{i=1}^{m} S_i \qquad (10)$$

where

$$S = -\int \ldots . \int p(x_1, x_2, \ldots, x_m) \log \; p(x_1, x_2, \ldots, x_m)\, dx_1 dx_2 \ldots dx_m$$

and

$$S_i = -\int p(x_i) \log p(x_i) dx_i \quad i = 1, 2, \ldots, m \qquad (11)$$

D_m can be used to measure dependency among genes. D_m is nonnegative i.e. $D_m >= 0$. If genes $x_1, x_2, x_3, \ldots, x_m$ are independent then $D_m = 0$, whereas if there is a dependency among the set of genes then $D_m > 0$. Higher value of D_m signifies greater dependency among the genes. For m dimensional multivariate normal density the values of joint entropy S and marginal entropy S_i [12] are given as follows:

$$S = \frac{m}{2} \log 2\pi + \frac{1}{2} \log \mid \Sigma \mid + \frac{1}{2} m \qquad (12)$$

$$S_i = \frac{1}{2} \log 2\pi + \frac{1}{2} \log {\sigma_i}^2 + \frac{1}{2} \quad i = 1, 2, \ldots, m \qquad (13)$$

Using equ. (12) and equ. (13), we can rewrite(4)equ. (10) as

$$D_m = \frac{1}{2} \sum_{i=1}^{m} {\sigma_i}^2 - \frac{1}{2} \log \mid \Sigma \mid \qquad (14)$$

Since the value of D_m increases with number of features so there is need of defining normalized value of D_m. It is known that marginal entropy S_i is always less than equal to joint entropy S i.e. $S_i \le S$ $i = 1, 2, \ldots, m$. This allows us to write

$$S_1 + S_2 + \ldots + S_m \le mS \qquad (15)$$

Using equ. (10) and equ. (15), we have

$$D_m = S_1 + S_2 + S_3 + \ldots . + S_m - S \le (m-1)S \qquad (16)$$

The normalized value of $\overline{D_m}$ is given by

$$\overline{D_m} = \frac{S_1 + S_2 + \ldots + S_m - S}{(m-1)S} \qquad (17)$$

Here, $0 \le \overline{D_m} \le 1$. Zero value of $\overline{D_m}$ corresponds to gene set consisting of independent genes. Higher value of $\overline{D_m}$ signifies more dependence among genes. To consider a set of independent genes, we can choose a threshold T. If the value of $\overline{D_m}$ is less than equal to threshold value T then gene subset is considered as a set of independent genes.

We have employed MI to measure relevance between a subset of genes and class label. Cross entropy is used to measure redundancy among genes. Our proposed algorithm MICE is incremental in nature and consists of two phase. In the first phase a set S of relevant and independent genes is created. We initially start with S as empty set and F a set which contain all the genes. Mutual information of each gene with respect to target class is estimated using equ 4. The gene which maximizes mutual information is selected. Let it be x_k, then S={x_k }and F=F- {$x_{k.}$}. Now a set of genes independent of selected gene subset S is created. This is done by calculating the cross entropy of SU{x_j} for all x_j in F. i.e. D(S,x_j). All the features whose D(S,x_j) >T are considered dependent with respect to the geneset S and hence these genes are removed from F. Again, a gene x_i in F which maximizes the mutual information of set S=SU{x_i} with respect to target class, is selected and included in set S. This gene x_i is removed from F. We determine consequently gene subset F which contains genes independent of S using cross entropy. This process is repeated till F becomes empty. In this way we create a set of independent and relevant genes.

In the second stage an optimal set of genes is determined from the gene subset selected in the first stage. To obtain the optimal set we have used a wrapper based forward feature selection. We have used classification accuracy as a criterion in the forward feature selection. The gene subset that maximizes the classification accuracy is selected. The outline of the proposed algorithm MICE is as follows.

MICE Algorithm

Input–Initial Set of genes,Class Labels C,Classifier M

```
PHASE 1 // to determine a subset of relevant and independent genes S
1. Intialization: Set F="initial set of genes" ; S = Φ //Set of Selected Attributes
2. Choose Threshold value T.
3. For each gene $x_i$ in F calculate $I(x_i;C)$ using (4)
4. Select the gene $x_k$ which maximizes Mutual Information $I(x_i;C)$ i.e. $x_k = \max_i I(x_i,C)$
5. $S = S \cup \{x_k\}$ ; $F = F - \{x_k\}$
6. Calculate $\overline{D}_m(S,x_j)$ for all $x_j \in F$ ;
   if $\overline{D}_m(S,x_j)$ >T   $F = F - \{x_j\}$ //Identifying set of independent genes $F$ with respect to S
7. Choose a gene from $x_k \in F$ which maximizes $I(S,x_k;C)$
8. $S = S \cup \{x_k\}$ , $F = F - \{x_k\}$
9. Repeat steps 6-8 till $F$ becomes empty
10. Return S

PHASE 2 // to determine subset of genes which provides maximum classification accuracy
1.  Initialization R = Φ
2.  For each $x_i \in S$ calculate classification accuracy for classifier M.
3.  $[x_k, \max\_acc] = \max_i Classif\_Acc(x_i)$
4.  $R = R \cup \{x_k\}$ ; $S = S - \{x_k\}$ ;   $R\_\min = R$
    // R_min is the gene subset corresponding to maximum accuracy
5.  For each $x_j \in S$ calculate classification_accuracy of S U{$x_j$} for classifier M
6.  $[x_k, new\_\max\_acc] = \max_i Classif\_Acc(S \cup x_i)$
7.  $R = R \cup \{x_k\}, S = S - \{x_k\}$
8.  If new_max_acc > max_acc  then R_min=R; max_acc=new_max_acc;
9.  Repeat steps 5-8 until max_acc=100 or S = Φ
10. Return R_min, max_acc
```

4 Experimental Setup and Results

To test the efficacy of our proposed algorithm MICE, we have carried out experiments on seven well known datasets. Colon, Leukemia, Prostate, Lung cancer and Ovary datasets are downloaded from Kent Ridge Bio-medical Dataset data repository [31]. For SRBCT we have used the dataset used by Khan [13]. NCI60 data is downloaded from the NCI Genomics and Bioinformatics Group Datasets resource [32]. The details of these datasets are given in Table 1. Before carrying experiments datasets are normalized using Z-score. In NCI60 dataset one class contained only two samples, so this class is removed from the dataset. Also number of samples belonging to each class is very small, therefore 2000 genes with highest variance are selected and then algorithm is applied on the reduced set of 2000 genes.

Table 1: Datasets Used.

Dataset	No. of Samples	No. of Features	Classes
Colon	62	2000	2
SRBCT	83	2308	4
Leukemia	72	7129	3
Prostate	102	5966	2
Ovary	253	15154	2
Lungcancer	181	12533	2
NCI60	60	2000	9

After normalizing the datasets, the first phase of our proposed algorithm MICE is applied to obtain the subset of relevant and independent genes. We performed experiments with different values of threshold. The value of threshold is varied from 0.1 to 0.9 with an increment of 0.1. It is observed that subset of genes selected is same for threshold values between 0.4 and 0.6. So, the value of threshold T is set as 0.5 in our experiments i.e. all the genes with $\overline{D}_m$ greater than 0.5 are rejected as dependent genes. The number of the reduced genes obtained after phase I of MICE algorithm for each dataset is given in Table 2.

Table 2: Size of Reduced Dataset after Phase I.

Dataset	Original No. Of genes	No. of genes selected
Colon	2000	56
SRBCT	2308	46
Leukemia	7129	54
Prostate	5966	66
Ovary	15154	98
Lungcancer	12533	159
NCI60	2000	60

It can be observed from Table 2 that the number of relevant genes obtained is significantly smaller in comparison to the original gene set. Now once a reduced set of relevant and independent genes is obtained, phase II of MICE algorithm is applied. Phase II of our proposed algorithm uses a forward feature selection strategy with a known classifier to obtain a set of genes which maximizes classification accuracy. We have employed four classifiers: linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), k-nearest neighbor (KNN) and support vector machine (SVM). Classification accuracy is calculated using leave-one-out cross validation (LOOCV). The algorithm is implemented in matlab. For KNN the optimal value of k is chosen. In SVM linear kernel is used. Results of our algorithm MICE are presented in Table 3. It contains maximum classification accuracy achieved along with the number of genes obtained by our algorithm MICE. It can be observed from Table 3 that our algorithm MICE is able to achieve maximum classification accuracy with small number of genes. For all the classifiers used we are able to achieve good classification accuracy with few numbers of genes. We compared the performance of our algorithm MICE with well known algorithm mRMR given by Peng et al (2005). The code of mRMR_d and mRMR_q is taken from [30]. As a preprocessing step, datasets are discretized into 3 values using $\mu \pm \sigma$. The values which are less than $\mu - \sigma$ are assigned -1, values between $\mu - \sigma$ and $\mu + \sigma$ are assigned value 0 and rest 1. Discretized data are passed to mRMR_d and mRMR_q and a ranked list of features is obtained from both the methods. Using the ranked list of genes obtained from mRMR, classification accuracy (LOOCV) is calculated as genes are added one by one. The maximum classification accuracy along with the minimum number of genes obtained for each classifier is shown in Table 3. We can observe the following from Table 3:

1. For Colon dataset a maximum accuracy of 96.77% is achieved with genes selected by our proposed algorithm MICE. It is achieved with 14 genes with QDC. For KNN results of MICE are better than mRMR. For SVM and LDC classification accuracy using MICE is same as mRMR.
2. For SRBCT dataset maximum classification accuracy of 100% is achieved with 15 genes with LDC and SVM classifier using MICE. For QDC results of MICE are better than mRMR. Only for KNN performance of mRMR is better.

Table 3: Comparison of maximum classification accuracy along with number of genes for different classifiers using various genes selection methods.

Dataset	Classifier	MICE (LOOCV)	mRMR_d (LOOCV)	mRMR_q (LOOCV)
Colon	LDC	**91.94(9)**	**91.94(3)**	90.32(6)
	QDC	**96.77(14)**	88.71(6)	87.10(27)
	KNN	**96.77(31)**	93.55(5)	91.94(32)
	SVM	**93.55(23)**	**93.55(18)**	93.55(50)
SRBCT	LDC	**100(15)**	97.59(21)	98.80(30)
	QDC	**98.80(10)**	49.40(66)	71.08(73)
	KNN	98.80(15)	100(83)	**100(29)**
	SVM	**100(15)**	100(19)	**100(15)**
Leukemia	LDC	**98.61(12)**	97.22(35)	98.61(14)
	QDC	**100(6)**	95.83(5)	88.89(9)
	KNN	**100(15)**	97.22(75)	97.22(80)
	SVM	98.61(7)	98.61(60)	**100(18)**
Prostate	LDC	**97.06(6)**	96.10(6)	96.08(8)
	QDC	**96.08(4)**	92.16(22)	89.22(9)
	KNN	**98.04(9)**	97.16(14)	98.04(27)
	SVM	**99.02(45)**	98.04(87)	98.04(26)
Ovary	LDC	**100(5)**	100(4)	100(8)
	QDC	**100(4)**	100(4)	100(8)
	KNN	**100(4)**	100(4)	100(10)
	SVM	**100(3)**	**100(5)**	100(8)
LungCancer	LDC	**100(44)**	100(36)	99.45(14)
	QDC	**100(5)**	100(40)	100(41)
	KNN	**100(3)**	100(20)	100(5)
	SVM	**100(4)**	100(23)	100(6)
NCI60	LDC	**84.48(26)**	75.86(94)	82.76(67)
	QDC	**56.90(5)**	56.90(5)	43.10(5)
	KNN	86.21(18)	**89.66(95)**	87.93(98)
	SVM	87.93(36)	81.03(34)	**89.66(97)**

3. For Leukemia dataset maximum classification accuracy of 100% is achieved with 6 genes in QDC classifier and 15 genes in KNN classifier using MICE. Same classification accuracy is achieved using MICE with LDC as with mRMR but with less number of genes.
4. For prostate dataset maximum classification accuracy of 99.02% is achieved with 45 genes in SVM classifier using MICE. Also for other classifiers, the accuracy achieved by our algorithm MICE is better in comparison to mRMR.
5. For Ovary dataset maximum classification accuracy of 100% is achieved for all classifiers using different gene selection methods but the number of genes selected by our proposed method MICE is same or comparitively less.
6. For Lungcancer dataset maximum classification accuracy of 100% is achieved for all classifiers using genes selected by our method MICE . The number of genes selected by MICE are significantly less in comparison to mRMR with QDC, KNN and SVM. The best result is obtained for KNN using only 3 genes.

   For NCI60 dataset maximum classification accuracy of 87.93% is achieved with 36 genes in SVM classifier using MICE. For LDC and QDC classifier accuracy achieved using genes selected by MICE is better in comparison to mRMR. For KNN and SVM performance of mRMR is better than MICE.
7. The performance of our proposed algorithm MICE with different classifiers is better in comparison to

mRMR algorithm in terms of classification accuracy in most cases. Our proposed algorithm MICE also provides a smaller subset of relevant genes for most of the cases.

The comparative results of classification accuracy obtained by different methods as the genes are added one by one for Leukemia dataset are shown in Figure 1. It can be observed from Figure 1 that classification accuracy obtained by our algorithm is more in comparison to mRMR_d and mRMR_q with the same number of genes for all the classifier. Similar results are observed for other datasets also.

To check the relevance of the selected genes subset we carried out 10 fold cross validation using the selected genes for all the datasets. Experiment is repeated 10 times. The average accuracy of 10 runs along with standard deviation is given in Table 4. It can be observed from the table that the 10 fold cross validation accuracy does not deviate much from LOOCV accuracy except for NCI60 dataset with QDC classifier. This shows that the gene set selected is not over fitted.

Table 4: 10 fold cross-validation accuracy achieved by the genes selected by MICE for different classifier. Quantity in bracket represents standard deviation.

Dataset	Classifier			
	LDC	QDC	KNN	SVM
Colon	91.13(1.14)	86.32(3.30)	**96.32(1.41)**	91.13(1.90)
SRBCT	99.40(0.85)	96.39(2.13)	98.07(1.3)	**99.52(1.02)**
Leukemia	96.94(1.71)	98.89(1.28)	**99.44(0.97)**	98.33(0.59)
Prostate	96.67(0.69)	96.20(1.08)	96.78(0.93)	**96.96(1.26)**
Ovary	**100(0)**	**100(0)**	99.88(0.27)	**100(0)**
LungC	99.17(0.39)	99.83(0.27)	99.89(0.23)	**100(0)**
NCI60	75.69(3.68)	41.72(2.12)	**81.72(2.72)**	79.66(4.36)

In literature a number of gene selection methods have been proposed and applied on these datasets. In Table 5, we have compared performance of our proposed method in terms of classification accuracy achieved and number of genes selected with some already existing gene selection methods in literature [6, 9, 10, 11, 13, 16, 18, 19, 20, 24, 25, 26, 27, 28, 29]. From Table 5, it can be observed that the performance of our proposed algorithm MICE is significantly better in terms of both classification accuracy and number of genes selected.

5 Conclusion

In this paper, we proposed a two stage algorithm MICE for finding a small subset of relevant genes responsible for better classification of high dimensional microarray datasets. The proposed method is based on the principle of Mutual Information and Cross Entropy. In first stage of algorithm, Mutual information is employed to select a set of relevant genes and Cross Entropy is used to determine independent genes. This provides a set of independent and relevant genes and reduces the size of gene set significantly. This allows us to use wrapper approach at the second stage. The use of wrapper method at the second stage gives a better subset of genes. Experimental results show that our proposed method MICE is able to achieve a better classification accuracy with small number of genes. In case of Lungcancer and Ovary 100% accuracy is achieved with 3 genes. For other datasets, the method provides competitive accuracy. Comparisons with other state-of-art methods show that our proposed algorithm is able to achieve better or comparable accuracy with less number of features in all the datasets.

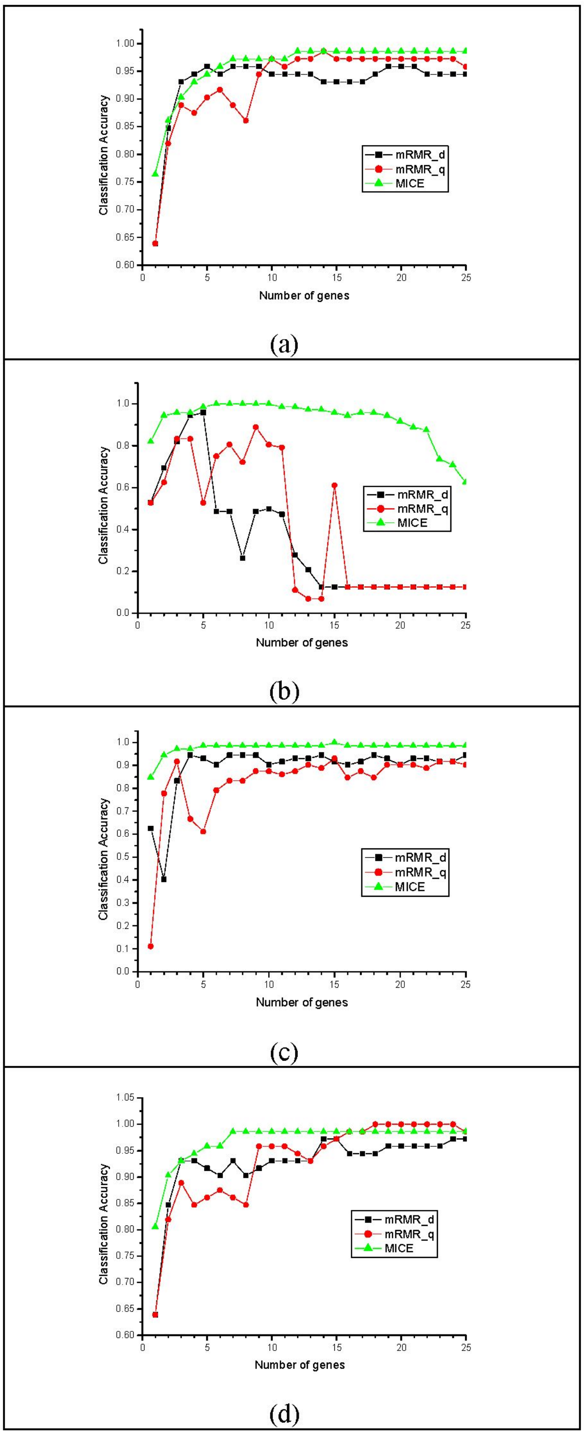

Figure 1: Classification accuracy Vs number of genes for Leukemia dataset using (a) LDC (b) QDC (c) KNN (d) SVM.

Table 5: Comparison of Maximum Classification accuracy and number of genes selected with other state of art methods.

COLON	
Proposed method	**96.77(14)**
PSO+ANN [27]	88.7
Chen and Zhao [29]	95.2
BIRSW [24]	85.48(3.50)
BIRSF [24]	85.48(7.40)
OVARY	
Proposed Method	**100(3)**
PSO+ANN [27]	97.0
NB [10]	96.2
BKS [10]	97.0
DT[10]	97.8
Chen and Zhao [29]	99.6
PROSTATE	
Proposed Method	**99.02(45)**
GAKNN [16]	84.6(205)
BIRS [24]	91.2(3)
Hong and Cho [[10]	96.3(79)
NCI60	
Proposed Method	**87.93(36)**
Jirapech-Umpai [11]	76.23
Liu [19]	88.52
Ooi [20]	85.37
Lin [18]	87.80
ReliefF/SVM [28]	58.33(30)
mRMR/RelieF [28]	68.33(30)
LEUKEMIA	
Proposed Method	**100(6)**
GS2+KNN [27]	98.6(10)
GS1+SVM [27]	98.6(4)
Cho's+SVM [27]	98.6(80)
Ftest + SVM [27]	98.6(33)
Fu and Liu [6]	97.0(4)
Guyon [9]	100(8)
Tibsrani [26]	100(21)
Chen and Zhao [29]	98.6
SRBCT	
Proposed Method	**100(15)**
GS2+SVM [27]	100(96)
GS1+SVM [27]	98.8(34)
Cho's+SVM [27]	98.8(80)
Ftest + SVM [27]	100(78)
Fu and Liu [6]	100(19)
Tibsrani [26]	100(43)
Khan [13]	100(96)
LUNGCANCER	
Proposed Method	**100(3)**
GS2+KNN [27]	93.1(44)
GS1+SVM [27]	98.6(4)
Cho's+SVM [27]	98.6(80)
Ftest + SVM [27]	98.6(94)
Shah and Kaushik [25]	100(8)
PSO+ANN [27]	98.3
Chen and Zhao [29]	98.3
GAKNN [16]	95.6(325)
Hong and Cho [10]	99.4(135)

References

[1] Battiti R (1994), Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Network 5(4) pp 537-550.

[2] Bellman R (1961), Adaptive Control Processes: A Guided Tour, Princeton University Press.

[3] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini (200), Tissue classification with gene gene expression profiles, In Proceedings of the fourth annual international conference on Computational molecular biology, pp 54-64, ACM Press.

[4] Bhattacharya C, Grate LR, Rizki A, Radisky D, Molina FJ, Jordan MI, Bissell MJ and Mian IS 92003), Simultaneously classification and relevant feature identification in high dimensional spaces: application to molecular profiling data, Signal Processing 83(4).

[5] Bo T and Jonassen I (2002), New feature subset selection procedures for classification of expression profiles, Genome biology 3.

[6] Fu LM and Liu CSF (2005), Evaluation of gene importance in microarray data based upon probability of selection, BMC Bioinformatics 6(67).

[7] Golub TR, Slonim DK, Tamayo, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri, Bloomfield CD and Lander ES (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 pp 531-537.

[8] Guyon I and Elisseeff A (2003), An Introduction to Variable and feature Selection, Journal of Machine Learning Research (3):1157-1182.

[9] Guyon I, Weston J, Barnhill S, Vapnik V (2003), Gene Selection for cancer classification using support vector machine, Machine Learning (46) pp 263-268.

[10] Hong JH and Cho SB (2006), The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming, Artif. Intell. Med. 36 pp 43–58.

[11] Jirapech-Umpai T and Aitken S (2005), Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes, BMC Bioinformatics 6:148.

[12] Kapur JN, Kesavan HK (1992) Entropy Optimization Principles with Applications, Academic Press.

[13] Khan J, Wei S, Ringner M, Saal LH, Ladanyi M, Westermann F (2001), Classification and diagnosis prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med 7 pp 673-679.

[14] Kohavi R and John G (1997), Wrapper for feature subset selection, Artificial Intelligence (1-2) pp 273-324.

[15] Kwak N and Choi CH (2002), Input Feature Selection for classification problems, IEEE Trans. Neural Netw 3(1) pp 143-159.

[16] Li L,Weinberg CR, Darden TA, Pedersen LG (2001), Gene Selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics 17(12) pp 1131-1142.

[17] Li T ,Zhang C, Ogihara M (2004), Comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics (20) pp 2429-2437.

[18] Lin TC, Liu RS, Chen CY, Chao YT and Chen SY (2006), Pattern classification in DNA microarray data of multiple tumor types, Pattern Recognition, 39 pp 2426-2438.

[19] Liu JJ, Cutler G, Li WX, Pan Z, Peng SH, Hoey T, Chen LB and Ling XFB (2005), Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 21 pp 2691-2697.

[20] Ooi CH, Tan P (2003), Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, Bioinformatics, 19 pp 37-44.

[21] Padmanabhan M and Dharanipragada S (2005), Maximizing Information Content in Feature Extraction, IEEE Transaction on speech and audio processing 13(4).

[22] Peng H, Long F, Ding C (2005), Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy, IEEE Trans. On Pattern Analysis and Machine Intelligence 27 pp 1226-1238.

[23] Ramaswamy S, Tamayo P (2001), Multiclass cancer diagnosis using tumour gene expression signature, Proc Natl Acad Sci, USA, 98(26) pp 15149-15154.

[24] Ruiz R, Riqueline J C, Aguilar-Ruiz JS (2006), Incremental wrapper based gene selection from microarray data for cancer classification, Pattern Recognition 39(12) pp 2383-2392.

[25] Shah, S and Kusiak A (2007), Cancer gene search with Data Mining and Genetic Algorithms, Computer in Biology medicine, Elsevier 37(2) pp 251-261.

[26] Tibsrani R , Hastie T, Narasimhan B and Chu G (2002), Diagnosis of multiple cancer types by shrunken centriods of gene expression, Proc. Natl Acad. Sci., USA (99) pp 6567-6572.

[27] Yang K, Cai Z, Li J, Lin G (2006), A stable gene selection in microarray data analysis, BMC Bioinformatics 7:228 .

[28] Yi Zhang, Chris HQ, Ding, Tao Li (2007), A Two-Stage Gene Selection Algorithm by Combining ReliefF and mRMR. BIBE 2007 pp 164-171.

[29] Chen Y and Zhao Y (2008), A novel ensemble of classifiers for microarray data classification, Applied Soft computing (8) pp 664-1669.

[30] http://www.mathworks.com/matlabcentral/fileexchange/14608.

[31] http://datam.i2r.a-star.edu.sg/datasets/krbd/

[32] http://discover.nci.nih.gov/datasetsNature2000.jsp